

third edition

Optoelectronics

an introduction

John Wilson
John Hawkes

Optoelectronics

$$M = 10^9 \text{ Hz}$$

Optoelectronics

An introduction

THIRD EDITION

JOHN WILSON

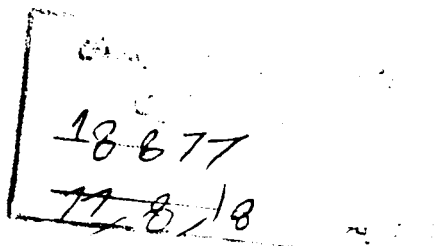
UNIVERSITY OF NORTHUMBRIA AT NEWCASTLE

JOHN HAWKES

UNIVERSITY OF NORTHUMBRIA AT NEWCASTLE



BL18877



PRENTICE HALL EUROPE

LONDON NEW YORK TORONTO SYDNEY TOKYO SINGAPORE

MADRID MEXICO CITY MUNICH PARIS



First published 1998 by
Prentice Hall Europe
Caprius 400, Maylands Avenue
Hemel Hempstead
Hertfordshire, HP2 7EZ
A division of
Simon & Schuster International Group

© Prentice Hall 1998

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form, or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission, in writing, from the publisher.

Typeset in 10/12pt Times
by Mathematical Composition Setters Ltd, Salisbury, Wiltshire

Printed and bound in Great Britain

Library of Congress Cataloging-in-Publication Data

Wilson, J. (John), 1939–

Optoelectronics : an introduction / John Wilson, John Hawkes. —
3rd ed.

p. cm.

Includes bibliographical references and index.

ISBN 0-13-103961-X

1. Optoelectronics. I. Hawkes, J. F. B., 1942– . II. Title.

QC673.W54 1998

621.381'045—dc21

97-36547

CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from
the British Library

ISBN 0-13-103961-X

1 2 3 4 5 02 01 00 99 98

Contents

<i>Preface to the third edition</i>	xi
<i>Glossary of symbols</i>	vii
1 Light	1
1.1 Nature of light	1
1.2 Wave nature of light	3
1.2.1 Polarization	7
1.2.2 Principle of superposition	11
1.2.3 Interference	14
1.2.4 Diffraction	19
1.3 Optical components	23
1.3.1 The spherical mirror	23
1.3.2 The thin spherical lens	24
1.3.3 Other lenses	26
1.4 Light sources – blackbody radiation	28
1.4.1 Blackbody sources	28
1.4.2 Line sources	30
1.5 Units of light	32
Notes	34
Problems	35
References	36
2 Elements of solid state physics	37
2.1 Review of some quantum mechanical concepts	37
2.1.1 Schrödinger equation	39
2.2 Energy bands in solids	42
2.2.1 Conductors, semiconductors and insulators	45
2.3 Electrical conductivity	48
2.4 Semiconductors	51
2.4.1 Intrinsic semiconductors	51
2.4.2 Extrinsic semiconductors	53
2.4.3 Excitons	56
2.5 Carrier concentrations	57

2.6	Work function	62
2.7	Excess carriers in semiconductors	63
2.7.1	Diffusion of carriers	64
2.7.2	Diffusion and drift of carriers	66
2.8	Junctions	66
2.8.1	The p-n junction in equilibrium	67
2.8.2	Current flow in a forward-biased p-n junction	70
2.8.3	Current flow in a reverse-biased p-n junction	73
2.8.4	Junction geometry and depletion layer capacitance	75
2.8.5	Deviations from simple theory	78
2.8.6	Other junctions	79
2.9	The quantum well	83
	Problems	86
	References	89
3	Modulation of light	90
3.1	Elliptical polarization	90
3.2	Birefringence	92
3.2.1	Phase plates	95
3.3	Optical activity	96
3.4	Electro-optic effect	96
3.4.1	Materials	106
3.5	Kerr modulators	107
3.5.1	Optical frequency Kerr effect	108
3.6	Scanning and switching	108
3.7	Magneto-optic devices	110
3.7.1	Faraday effect	110
3.8	Acousto-optic effect	112
3.9	Quantum well modulators	117
3.10	Non-linear optics	119
3.10.1	Parametric oscillation	124
	Problems	126
	References	127
4	Display devices	129
4.1	Luminescence	129
4.2	Photoluminescence	131
4.3	Cathodoluminescence	133
4.4	Cathode ray tube	134
4.5	Electroluminescence	138
4.6	Injection luminescence and the light-emitting diode	141
4.6.1	Radiative recombination processes	142
4.6.2	LED materials	146
4.6.3	Commercial LED materials	147

4.6.4 LED construction	149
4.6.5 Response times of LEDs	152
4.6.6 LED drive circuitry	153
4.7 Plasma displays	155
4.8 Display brightness	157
4.9 Liquid crystal displays	158
4.10 Numeric displays	163
Notes	166
Problems	167
References	167
5 Lasers I	169
5.1 Emission and absorption of radiation	169
5.2 Einstein relations	171
5.3 Absorption of radiation	173
5.4 Population inversion	175
5.4.1 Attainment of a population inversion	177
5.5 Optical feedback	179
5.6 Threshold conditions – laser losses	181
5.7 Lineshape function	183
5.8 Population inversion and pumping threshold conditions	186
5.9 Laser modes	190
5.9.1 Axial modes	190
5.9.2 Transverse modes	193
5.10 Classes of laser	195
5.10.1 Doped insulator lasers	196
5.10.2 Semiconductor lasers	204
5.10.3 Gas lasers	223
5.10.4 Liquid dye lasers	233
5.10.5 Parametric lasers	236
5.10.6 The free electron laser	238
5.11 Conclusion	239
Notes	240
Problems	240
References	241
6 Lasers II	244
6.1 Single mode operation	244
6.2 Frequency stabilization	245
6.3 Mode locking	250
6.3.1 Active mode locking	252
6.3.2 Passive mode locking	253
6.4 Q-switching	254
6.4.1 Methods of Q-switching	255

6.5	Laser applications	258
6.5.1	Properties of laser light	258
6.6	Measurement of distance	267
6.6.1	Interferometric methods	267
6.6.2	Beam modulation telemetry	269
6.6.3	Pulse echo techniques	271
6.7	Holography	271
6.7.1	Applications of holography	274
6.7.2	Holographic computer memories	276
6.8	High energy applications of lasers	278
6.8.1	Industrial applications	281
6.8.2	Medical applications	285
6.8.3	Laser-induced nuclear fusion	286
	Problems	288
	References	289
7	Photodetectors	293
7.1	Detector performance parameters	293
7.2	Thermal detectors	296
7.2.1	Thermoelectric detectors	298
7.2.2	The bolometer	298
7.2.3	Pneumatic detectors	300
7.2.4	Pyroelectric detectors	300
7.3	Photon devices	303
7.3.1	Photoemissive devices	303
7.3.2	Vacuum photodiodes	305
7.3.3	Photomultipliers	307
7.3.4	Image intensifiers	312
7.3.5	Photoconductive detector	314
7.3.6	Junction detectors	324
7.3.7	Detector arrays	344
7.3.8	Liquid crystal light valves	348
7.3.9	Photon counting techniques	352
7.3.10	Solar cells	353
	Notes	355
	Problems	355
	References	358
8	Fiber optical waveguides	359
8.1	Total internal reflection	360
8.2	Planar dielectric waveguide	364
8.3	Optical fiber waveguides	373
8.3.1	Step index multimode fibers	376
8.3.2	Intermodal dispersion	377
8.3.3	Graded index fibers	382

8.3.4 Single mode fibers	385
8.3.5 Fiber materials and types	388
8.3.6 Dispersion in single mode fibers	389
8.4 Losses in fibers	393
8.4.1 Bending losses	393
8.4.2 Intrinsic fiber losses	394
8.5 Optical fiber connectors	397
8.5.1 Single fiber jointing	397
8.5.2 Fiber couplers	402
8.6 Measurement of fiber characteristics	406
8.6.1 Introduction	406
8.6.2 Fiber attenuation measurements	407
8.6.3 Fiber dispersion measurements	408
8.6.4 Cut-off wavelengths in single mode fiber	409
8.6.5 Refractive index profile measurement	410
8.6.6 Optical time domain reflectometer	411
8.7 Fiber materials and manufacture	413
8.7.1 Silica-based fibers	413
8.7.2 Plastic-coated silica fiber	414
8.7.3 All-plastic fibers	415
8.7.4 Mid-infrared fibers	417
8.7.5 Special fiber types	418
8.8 Fiber cables	421
Notes	423
Problems	423
References	426

9 Optical communication systems 428

9.1 Modulation schemes	428
9.1.1 Analog modulation	429
9.1.2 Digital modulation	432
9.2 Free space communications	436
9.3 Fiber optical communication systems	438
9.3.1 Operating wavelength	440
9.3.2 Emitter design	440
9.3.3 Detector design	448
9.3.4 Fiber choice	457
9.3.5 Optical amplifiers	457
9.3.6 System design considerations	461
9.3.7 Local area networks	465
9.3.8 Wavelength division multiplexing	466
9.3.9 Coherent systems	468
9.3.10 Solitons	470
9.4 Integrated optics	472
9.4.1 Slab and stripe waveguides	472
9.4.2 Basic IO structural elements	475

9.4.3 IO devices	484
Notes	487
Problems	487
References	490

10 Non-communications applications of fibers 492

10.1 Optical fiber sensors	492
10.1.1 Multimode extrinsic optical fiber sensors	492
10.1.2 Multimode intrinsic optical fiber sensors	497
10.1.3 Distributed fiber sensors	500
10.1.4 Single mode fiber sensors	503
10.2 Light-guiding fibers	515
10.2.1 Coherent bundles	519
Problems	521
References	522

APPENDICES

1 Answers to numerical problems	524
2 Birefringence	527
3 Limitations on LED frequency response due to carrier diffusion and recombination	535
4 Interactions between radiation and electronic energy levels with finite frequency linewidths	537
5 Optical bandwidths and pulse broadening	541
6 Physical constants and properties of some common semiconductors at room temperature (300 K)	546
7 Laser safety	547

<i>Index</i>	551
--------------	-----

Preface to the third edition

In the six or seven years which have elapsed since the second edition was published the importance of optoelectronics as a subject in its own right has continued to grow, and the applications of optoelectronic devices have increased significantly. Very few courses in physics or electronic engineering do not now include a discussion of optoelectronics.

Our interpretation of what is meant by optoelectronics has remained unaltered since the first edition was published in 1983. That is, we define optoelectronics as the interaction of light (in the wavelength range from about 100 nm to 20 μm) with matter in gaseous, liquid or solid form, and the devices which depend on these interactions. This definition is of course broader than that adopted by many authors who restrict their discussions to the ways in which light interacts with semiconductors. This, however, ignores the many important devices that depend on the behaviour of light in crystals subject to external force fields, and the majority of lasers.

Typical of the dramatic growth of optoelectronics is the staggering rise in fiber optic communications. Optical fibers with laser sources have enabled home subscribers to have access to an enormous amount of facilities and information, varying from telephone links, through many video channels, to information databases worldwide via the information superhighway. The so-called information superhighway has resulted from the development of very low dispersion fibers coupled with the enormous bandwidth provided by laser sources, and the availability of very fast light detectors. Similarly there has been an amazing continuation in the growth of laser applications, not least in the field of medicine.

The book was originally written very much with the final year UK undergraduate in mind. It has, however, been widely used as an introduction to optoelectronics for postgraduate students and those in industry, who require a treatment that is not too advanced, but which nevertheless gives a good introduction to the quantitative aspects of the subject. We see no reason to change this approach as the book has been used as a standard text by colleagues in very many institutions worldwide for both undergraduate and postgraduate students. We are grateful to those who, having used the book, have taken the time to point out minor errors and to make suggestions for improvements.

This edition then does not aim to cover all aspects of optoelectronics nor to deal with some topics in full theoretical rigour, which in many cases would require a formal quantum mechanical approach. It aims rather to put special emphasis on the fundamental principles which underlie the operation of devices and systems. This, it is anticipated, will enable the reader to appreciate the operation of devices not covered here and to understand future developments within the subject.

Optoelectronics relies heavily on the disciplines of optics and solid state physics and we expect the reader to have some background knowledge of these subjects. For those with little experience in these topics we have retained, after careful consideration, the original Chapters 1 and 2 with some modifications, which provide a review of some of the relevant topics in these areas.

Otherwise in this third edition, given the rapid developments mentioned earlier, we have taken the opportunity to update the material covered in all of the other chapters. We have also reduced the length of some sections and introduced new sections and topics to reflect the changing emphasis within the subject. Thus in Chapter 3 we have increased the emphasis on parametric oscillation, and introduced a section on solitons in Chapter 9.

The two chapters on lasers, Chapters 5 and 6, have been updated to include, for example, sections on the free electron laser, quantum well lasers, vertical emitting lasers and superluminescent diodes, fiber lasers and parametric lasers, together with a section on medical and industrial applications of lasers.

In Chapter 7 there is a reduced emphasis on thermal detectors but an increased emphasis on junction detectors, especially those operating in the near IR and on high speed devices.

Chapter 8 on optical fibers has been updated to emphasize the increased importance of single mode fibers and very low dispersion (high bandwidth) fibers. Fiber manufacture is updated and the production of long wavelength fibers discussed.

In Chapter 9 wavelength division multiplexing, optical amplifiers, solitons and coherent systems are all introduced, along with a consideration of the performance of systems which depend on these topics. The systems covered include local area networks (LANs) and world-wide telephone links. The section on integrated optoelectronics has been updated and expanded.

Finally, while the optical fiber sensor which rivals the ubiquitous thermocouple does not yet appear to have been developed, there is a continued high level of interest in optical fiber sensors and some recent developments, particularly on distributed systems, have been included in Chapter 10.

Many readers of the previous editions have indicated the usefulness of the in-chapter examples and we have retained these and introduced several more. These, we believe, give students a feel for the subject and of the orders of magnitude of the parameters involved, and provide a better understanding of the text. We have also included more end-of-chapter problems. A teachers' manual containing solutions to these problems can be obtained from the publisher. The book again uses SI units throughout with the exception of the occasional use of the electron volt (eV).

Glossary of symbols

\mathcal{A}	Source strength
A	area, electric field amplitude, spontaneous transition rate (A_{21})
a	Richardson–Dushman constant, fiber radius, periodicity of lattice
B	magnetic flux density, Einstein coefficient (B_{21} , B_{12}), electron–hole recombination parameter, luminance, ‘flicker’ noise constant, birefringence
BER	bit error rate
C	capacitance, waveguide coupling factor
D	diffusion coefficient (D_e , D_h)
d	mode volume thickness
\mathcal{E}	electric field
E	energy, bandgap (E_g), donor/acceptor energy level (E_d , E_a), phonon energy (E_p), exciton binding energy (E_c), Young’s modulus
F	fractional transmission, lens f number, force, APD excess noise factor ($F(M)$), Fermi–Dirac distribution function ($F(E)$), solar cell fill factor, electric field decay factor ($F(y)$)
f	modulation frequency, cut-off frequency (f_c), focal length
G	thermal conductance, gain
g	degeneracy, electron–hole generation rate, lineshape function ($g(v)$)
\mathcal{H}	magnetic field
H	heat capacity, system frequency response ($H(f)$)
h	polarization holding parameter, normalized impulse response ($h(f)$)
h_{fe}	transistor common emitter current gain
\mathcal{I}	radiant or luminous intensity
I	irradiance
i	current, reverse bias saturation current (i_0), photoinduced current (i_λ)
i	$\sqrt{-1}$
\hat{i}	unit vector (x direction)
\mathcal{J}	molecular rotational quantum number
J	current density
\hat{j}	unit vector (y direction)
\mathcal{K}	diffraction factor
K	Kerr constant, electron beam range parameter
k, k	wavevector, wavenumber, small signal gain coefficient
\mathcal{L}	inductance

L	diffusion length (L_c, L_{th}), beat length (L_p), radiance, insertion loss (L_{in}), excess loss (L_e)
l_c	coherence length
M	mass, avalanche multiplication factor
m	mass, effective mass (m_e^*, m_h^*), image magnification
N	number of photons
N	population inversion, donor/acceptor densities (N_d, N_a), effective density of states in conduction/valence band (N_c, N_v), number of photons (N_p), number of modes, group refractive index (N_g)
NA	numerical aperture
NEP	noise equivalent power
n	electron concentration, intrinsic carrier concentration (n_i), refractive index, quantum number, mode number
O_d	detector output
ϕ	phase factor
P	power, dipole moment, electrical polarization, quadratic electro-optic coefficient
p	hole concentration, momentum, probability, photoelastic coefficient (p_e)
Q	charge, 'quality factor', trap escape factor, profile dispersion parameter, radiant or luminous energy
R	electrical resistance, load resistor (R_L), radius of curvature, reflectance, frequency response ($R(f)$), Stokes to anti-Stokes scattering ratio, responsivity, electron range (R_e), Fresnel reflection loss (R_F)
r	linear electro-optic coefficient, ratio of electron to hole ionization probabilities, electron-hole generation/recombination rates (r_g, r_r), reflection coefficient
S_R	Rayleigh scattering fraction
S/N	signal-to-noise ratio
T	transmittance, temperature, Curie temperature (T_c), period
t	time, active region thickness
$U_0(x, y)$	electric field amplitude
V	fringe visibility
V	voltage, potential energy, Verdet constant, normalized film thickness, eye relative spectral response
v	velocity, group velocity (v_g), molecular vibration quantum number, Poisson ratio
W	power, total depletion layer width, spectral radiant emittance
$x_{n,p}, x$	n, p depletion layer widths, coordinate distance
y	coordinate distance
Z	depth of field, density of states ($Z(E)$)
z	coordinate distance
α	absorption coefficient, temperature coefficient of resistance, transistor common base current gain, angle, fiber profile parameter
β	diode ideality factor, electron-hole generation efficiency factor, propagation constant, isothermal compressibility, refractive index temperature coefficient
γ	loss coefficient, mutual coherence function (γ_{12})
Δ	fiber refractive index ratio

Δt	coherence time
δ	phase angle, secondary electron emission coefficient, waveguide difference parameter
ϵ	relative permittivity/dielectric constant (ϵ_r), emissivity
η	efficiency, charge transfer efficiency (η_{ct})
θ, θ_B	angle, Brewster angle
Λ	acoustic wavelength, microbend wavelength, grating periodicity
λ	light wavelength, bandgap wavelength (λ_g), light wavelength in vacuum (λ_0)
μ	electron/hole mobility (μ_e, μ_h), relative permeability (μ_r)
ν	light wave frequency
ρ	charge density, radiation density (ρ_r), resistivity
σ	conductivity, Stefan's constant, r.m.s. pulse width
τ	time constant, thermal time constant (τ_H), lifetime, minority carrier lifetime (τ_c), time
Φ	phase angle, light flux
ϕ	phase angle, work function
χ	electric susceptibility, electron affinity
Ψ	time-dependent wavefunction
ψ	time-independent wavefunction, phase change
Ω	solid angle, rotation rate
ω	angular frequency, mode field diameter (ω_0)

Light

In discussing the various topics which we have brought together in this text under the title *optoelectronics*, of necessity we rely heavily on the basic physics of light, matter and their interactions. In this and the next chapter we describe rather briefly those concepts of optics and solid state physics which are fundamental to optoelectronics. The reader may be familiar with much of the content of these two chapters though those who have not recently studied optics or solid state physics may find them useful. For a more detailed development of the topics included, the reader is referred to the many excellent texts on these subjects, a selection of which is given in refs 1.1 and 2. 1.

In this chapter we shall describe phenomena such as polarization, diffraction, interference and coherence; we have assumed that the basic ideas of the reflection and refraction of light and geometrical optics are completely familiar to the reader, though one or two results of geometrical optics are included for convenience. In this context it is worth noting that the term 'light' is taken to include the ultraviolet and near-infrared regions as well as the visible region of the spectrum.

1.1 Nature of light

During the seventeenth century two emission theories on the nature of light were developed, the wave theory of Hooke and Huygens and the corpuscular theory of Newton. Subsequent observations by Young, Malus, Euler and others lent support to the wave theory. Then in 1864 Maxwell combined the equations of electromagnetism in a general form and showed that they suggest the existence of transverse electromagnetic waves. The speed of propagation in free space of these waves was given by

$$c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} \quad (1.1)$$

where μ_0 and ϵ_0 are the permeability and permittivity of free space, respectively. Substitution of the experimentally determined values of μ_0 and ϵ_0 yielded a value for c in very close agreement with the value of the speed of light *in vacuo* measured independently. Maxwell therefore proposed that light was an electromagnetic wave having a speed c of approximately $3 \times 10^8 \text{ m s}^{-1}$, a frequency of some $5 \times 10^{14} \text{ Hz}$ and a wavelength of about 500 nm. Maxwell's theory suggested the possibility of producing electromagnetic waves

TABLE 1.1 Electromagnetic spectrum

Type of radiation	Wavelength	Frequency (Hz)	Quantum energy (eV)
Radio waves	100 km	3×10^3	1.2×10^{-31}
	300 nm	10^9	4×10^{-6}
Microwaves	0.3 mm	10^{12}	4×10^{-3}
	0.7 μm	4.3×10^{14}	1.8
Infrared	0.4 μm	7.5×10^{14}	3.1
	0.03 μm	10^{16}	40
Visible	0.1 nm	3×10^{18}	1.2×10^4
	1.0 pm	3×10^{20}	1.2×10^6
Ultraviolet			
X-rays			
γ rays			

Note: The divisions into the various regions are for illustration only; there is no firm dividing line between one region and the next. The numerical values are only approximate; the upper and lower limits are somewhat arbitrary.

with a wide range of frequencies (or wavelengths). In 1887 Hertz succeeded in generating non-visible electromagnetic waves, with a wavelength of the order of 10 m, by discharging an induction coil across a spark gap thereby setting up oscillating electric and magnetic fields. Visible light and Hertzian waves are part of the *electromagnetic spectrum* which, as we can see from Table 1.1, extends approximately over the wavelength range of 1.0 pm to 100 km. The wave theory thus became the accepted theory of light. However, while the wave theory, as we shall see below, provides an explanation of optical phenomena such as interference and diffraction, it fails completely when applied to situations where energy is exchanged, such as in the emission and absorption of light and the photoelectric effect. The photoelectric effect, which is the emission of electrons from the surfaces of solids when irradiated, was explained by Einstein in 1905. He suggested that the energy of a light beam is not spread evenly but is concentrated in certain regions, which propagate like particles. These 'particles' of energy subsequently became known as photons (G. N. Lewis, 1926).

Einstein was led to the concept of photons by the work of Planck on the emission of light from hot bodies. Planck found that the observations indicated that light energy is emitted in multiples of a certain minimum energy unit. The size of the unit, which is called a *quantum*, depends on the frequency ν of the radiation and is given by

$$E = h\nu \quad (1.2)$$

where h is Planck's constant. Planck's hypothesis did not require that the energy should be emitted in *localized* bundles and it could, with difficulty, be reconciled with the electromagnetic wave theory. When Einstein showed, however, that it seemed necessary to assume

the concentration of energy travelling through space as particles, a wave solution was excluded. Thus we have a particle theory also; light apparently has a *dual* nature!

The two theories of light are not in conflict but rather they are complementary. For our purposes it is sufficient to accept that in many experiments, especially those involving the exchange of energy, the particle (photon or quantum) nature of light dominates the wave nature. On the other hand, for experiments involving interference or diffraction, where light interacts with light, the wave nature dominates.

1.2

Wave nature of light

Light as an electromagnetic wave is characterized by a combination of time-varying electric (\mathcal{E}) and magnetic (\mathcal{H}) fields propagating through space (see e.g. ref. 1.1b, Chapters 19–21; 1.1c, Chapter 3). Maxwell showed that both these fields satisfy the same partial differential equation:

$$\nabla^2(\mathcal{E}, \mathcal{H}) = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} (\mathcal{E}, \mathcal{H}) \quad (1.3)$$

This is called the *wave equation*; it is encountered in many different kinds of physical phenomena such as mechanical vibrations of a string or in a rod. The implication of eq. (1.3) is that *changes* in the fields propagate through space with a speed c , the speed of light. The frequency of oscillation of the fields, ν , and their wavelength in vacuum, λ_0 , are related by

$$c = \nu \lambda_0 \quad (1.4)$$

In any other medium the speed of propagation is given by

$$\nu = \frac{c}{n} = \nu \lambda = \nu \frac{\lambda_0}{n} \quad (1.5)$$

where n is the refractive index of the medium and λ is the wavelength in the medium (later in the text we often drop the subscript from the vacuum wavelength λ_0 to simplify the notation). n is given by

$$n = \sqrt{\mu_r \epsilon_r} \quad (1.5a)$$

where μ_r and ϵ_r are the relative permeability and relative permittivity of the medium respectively.

The electric and magnetic fields vibrate perpendicularly to one another and perpendicularly to the direction of propagation as illustrated in Fig. 1.1; that is, light waves are transverse waves. In describing optical phenomena we often omit the magnetic field vector. This simplifies diagrams and mathematical descriptions but we should always remember that there is also a magnetic field component which behaves in a similar way to the electric field component.

The simplest waves are sinusoidal waves, which can be expressed mathematically by the equation

$$\mathcal{E}(x, t) = \mathcal{E}_0 \cos(\omega t - kx + \phi) \quad (1.6)$$

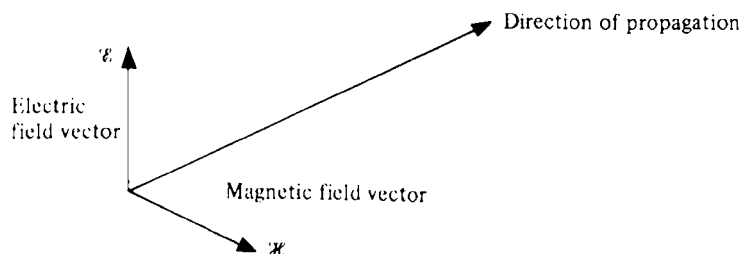


FIG. 1.1 Electromagnetic wave: the electric vector (\mathcal{E}) and the magnetic vector (\mathcal{H}) vibrate in orthogonal planes and perpendicularly to the direction of propagation.

where \mathcal{E} is the value of the electric field at the point x at time t , \mathcal{E}_0 is the amplitude of the wave, ω is the angular frequency ($\omega = 2\pi\nu$), k is the wavenumber ($k = 2\pi/\lambda$) and ϕ is the phase constant. The term $(\omega t - kx + \phi)$ is the phase of the wave. Equation (1.6), which describes a perfectly monochromatic plane wave of infinite extent propagating in the positive x direction, is a solution of the wave equation (1.3).

We can represent eq. (1.6) diagrammatically by plotting \mathcal{E} as a function of either x or t as shown in Figs 1.2(a) and (b), where we have taken $\mathcal{E} = \mathcal{E}_0$ at x and t equal to zero so that $\phi = 0$. Figure 1.2(a) shows the variation of the electric field with distance at a given instant of time. If, as a representative time, we take t equal to zero, then the spatial variation of the electric field is given by

$$\mathcal{E} = \mathcal{E}_0 \cos kx \quad (1.6a)$$

Similarly Fig. 1.2(b) shows the variation of the electric field as a function of time at some specific location in space. If we take x equal to zero then the temporal variation of electric field is given by

$$\mathcal{E} = \mathcal{E}_0 \cos \omega t \quad (1.6b)$$

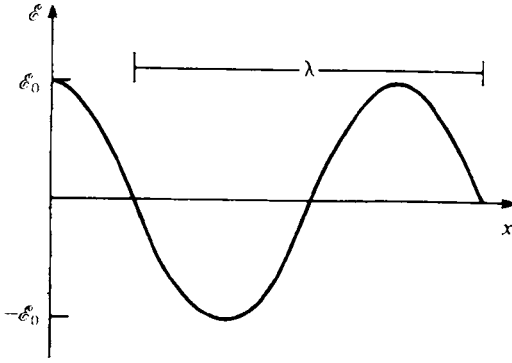
Equations (1.6) can be written in a variety of equivalent forms using the relationships between ν , ω , λ , k and c already given. We note also that the time for one cycle is the period T ($T = 1/\nu$) as shown in Fig. 1.2(b).

If the value of \mathcal{E} at $x = 0$, $t = 0$ is not \mathcal{E}_0 then we must include the arbitrary phase constant ϕ . Equations (1.6) can also be expressed using a sine rather than a cosine function, or alternatively using complex exponentials.

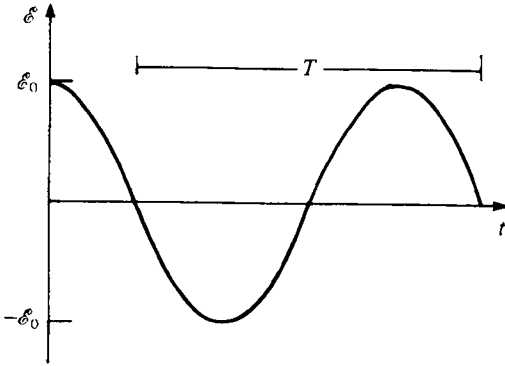
In the plane waves described above and in other forms of wave there are surfaces of constant phase, which are referred to as wave surfaces or wavefronts. As time elapses the wavefronts move through space with a velocity v given by

$$v = \omega/k = \nu\lambda \quad (1.7)$$

which is called the *phase* velocity. As it is impossible in practice to produce perfectly monochromatic waves we often have the situation where a group of waves of closely similar wavelength is moving such that their resultant forms a packet. This packet moves with the *group* velocity v_g . A discussion of this phenomenon based on the combination of two waves of



(a)



(b)

FIG. 1.2 Electric field (\mathcal{E}) of an electromagnetic wave plotted as a function of (a) the spatial coordinate x and (b) the time t .

slightly different frequencies moving together, which is illustrated in Fig. 1.3, shows that the group velocity is given by (see Problem 1.2)

$$v_g = \frac{\partial \omega}{\partial k} \quad (1.8)$$

Equations (1.6) represent plane waves moving along the x axis; we can generalize our mathematical description to include plane waves moving in arbitrary directions. Such a wave can be characterized by a wavevector \mathbf{k} where $|\mathbf{k}| = 2\pi/\lambda$ and eq. (1.6) becomes

$$\mathcal{E}(x, y, z, t) = \mathcal{E}_0 \cos(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi) \quad (1.9)$$

where \mathbf{r} is a vector from the origin to the point (x, y, z) . Thus, for example, if we have a plane wave propagating in a direction θ to the x axis with its wavefronts normal to the xy plane

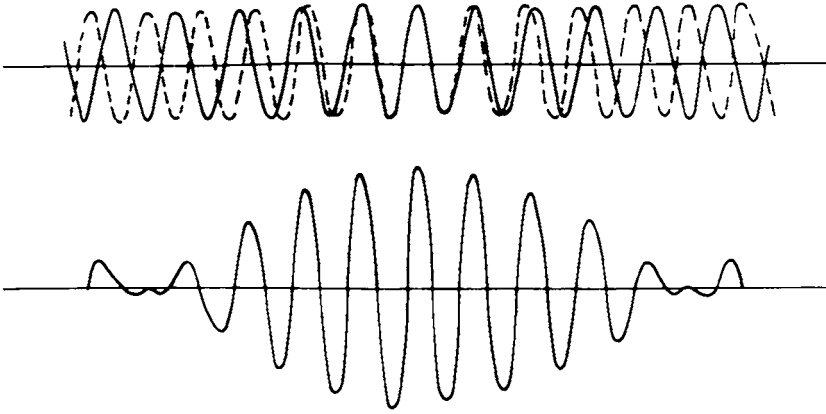


FIG. 1.3 The concept of a wave group or wave packet is illustrated by the combination of two progressive waves of nearly equal frequency. The envelope of the group moves with the group velocity v_g

as shown in Fig. 1.4, we can write

$$\mathbf{k} = ik_x + jk_y \quad (1.10)$$

and

$$\mathbf{r} = ix + jy \quad (1.11)$$

where i and j are unit vectors in the x and y directions respectively. Combining eqs (1.10)

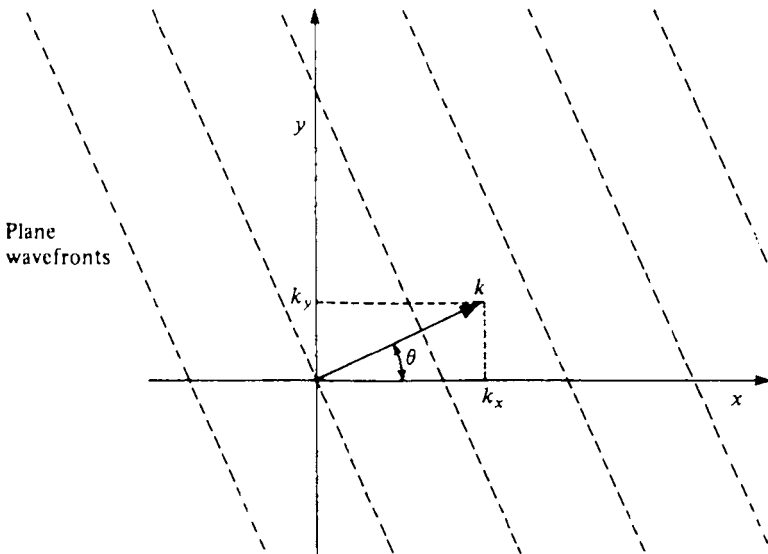


FIG. 1.4 Plane wave with its propagation vector \mathbf{k} in the xy plane. The components of the propagation vector are $k_x = |\mathbf{k}| \cos \theta$ and $k_y = |\mathbf{k}| \sin \theta$.

and (1.11) we have

$$\mathbf{k} \cdot \mathbf{r} = xk_x + yk_y = xk \cos \theta + yk \sin \theta$$

Hence we can write eq. (1.9) in this case as

$$\xi(x, y, t) = \xi_0 \cos(\omega t - xk \cos \theta - yk \sin \theta + \phi) \quad (1.9a)$$

An equally important concept is that of spherical waves which, we can imagine, are generated by a point source of light. If such a source is located in an isotropic medium it will radiate uniformly in all directions; the wavefronts are thus a series of concentric spherical shells. We can describe this situation by

$$\xi = \frac{\mathcal{A}}{r} \cos(\omega t - \mathbf{k} \cdot \mathbf{r})$$

where the constant \mathcal{A} is known as the *source strength*. The factor $1/r$ in the amplitude term accounts for the decrease in amplitude of the wave as it propagates further and further from the source. As the irradiance¹ is proportional to the square of the amplitude, there is an inverse-square-law decrease in irradiance. If the medium in which the source is located is anisotropic, then the wave surfaces are no longer spheres; their shapes depend on the speed of propagation in different directions. We shall return to this point in section 3.2.

The path followed by the normal to the wavefronts or wave surfaces is often described by the term *light rays*, and we may think of a beam of light as comprising a very large bundle of such light rays moving in the same direction. Similarly we can easily envisage light rays diverging uniformly outwards in all directions from a point source of light situated in a homogeneous region. The concept of light rays is useful when analyzing, to a first approximation, the behaviour of light as it passes through a 'system' of optical components such as lenses and mirrors – see section 1.3.

1.2.1 Polarization

If the electric field vector of an electromagnetic wave propagating in free space vibrates in a specific plane, the wave is said to be plane polarized. Any real beam of light comprises many individual waves and in general the planes of vibration of their electric fields will be randomly orientated. Such a beam of light is unpolarized and the resultant electric field vector changes orientation randomly in time. It is possible, however, to have light beams characterized by highly orientated electric fields and such light is referred to as being *polarized*. The simplest form of polarization is plane polarized light, which is similar to the single wave shown in Fig. 1.1. Other forms of polarization are discussed in section 3.1.

Light can be polarized in a number of different ways and here we consider two, namely polarization by reflection and by absorption. When unpolarized light is incident on a material surface as shown in Fig. 1.5(a) we find that the light with its polarization vector perpendicular to the plane of incidence (denoted by \longleftrightarrow) is preferentially reflected in comparison with light polarized parallel to the plane of incidence (denoted by $\uparrow\uparrow\uparrow$). We can resolve the electric field vector of each wave into components parallel and perpendicular to any convenient direction: here we choose the plane of incidence. Thus by symme-

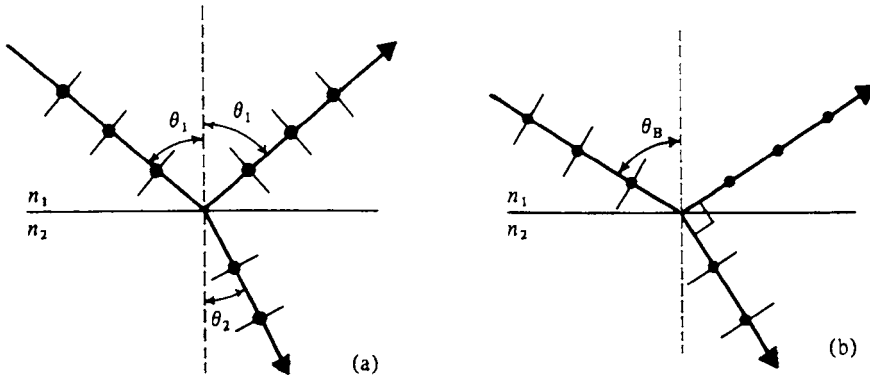


FIG. 1.5 In (a), a light reflected from the interface between two media of refractive index n_1 and n_2 is partially plane polarized (there is less of the parallel component $\perp\perp$ than the perpendicular component $\cdot\cdot\cdot$), while in (b) for incidence at the Brewster angle θ_B , the reflected light is completely plane polarized.

try we may think of unpolarized light as comprising two equal plane polarized components with orthogonal orientations.

We find further that the reflectance² of the surface for the perpendicular and parallel components varies as a function of the angle of incidence as shown in Fig. 1.6. In particular we note that for the parallel component the reflectance is zero at the specific angle of incidence $\theta = \theta_B$. For this angle of incidence, which is called the *Brewster angle*, all of the parallel component is transmitted. It is found also that for incidence at the Brewster angle the reflected and refracted (transmitted) rays are perpendicular to one another as illustrated in Fig. 1.5(b). Thus, using Snell's law for the refraction of light at the interface between two media of refractive indices n_1 and n_2 , that is

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (1.12)$$

where θ_1 and θ_2 are the angles of incidence and refraction respectively, we have

$$n_1 \sin \theta_B = n_2 \cos \theta_B$$

or

$$\tan \theta_B = \frac{n_2}{n_1} \quad (1.13)$$

Equation (1.13) is known as Brewster's law. One way of polarizing light is simply to pass it through a series of glass plates orientated at the Brewster angle. Then at each surface, some of the light polarized perpendicularly to the plane of incidence is reflected, as shown in Fig. 1.7, while all of the parallel component is transmitted. After passing through about six plates, the transmitted light is highly plane polarized. We shall see in Chapter 5 that lasers often include surfaces inclined at the Brewster angle to minimize optical losses for one particular orientation of polarization of the light passing through the surfaces. Such lasers thus emit plane polarized light.

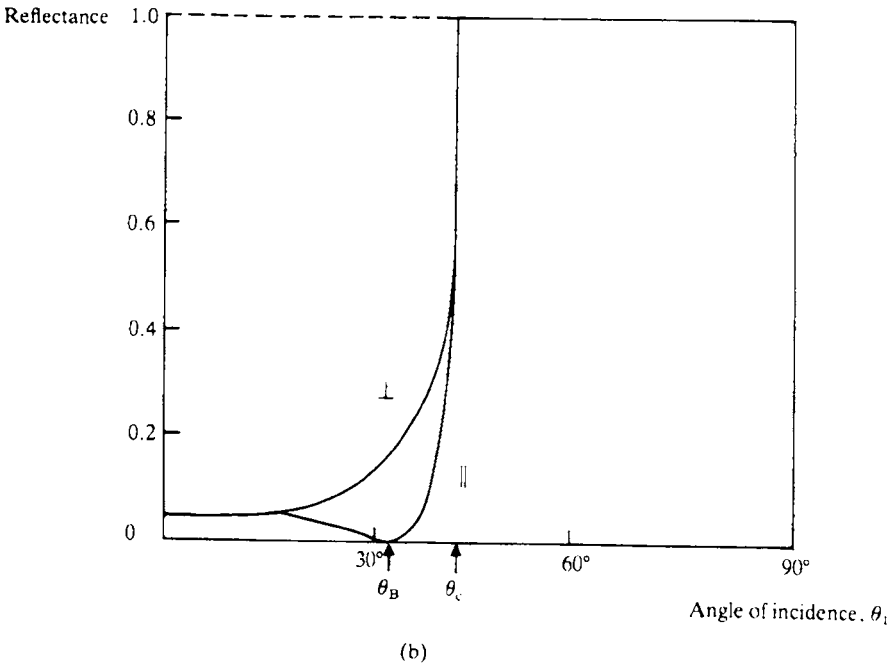
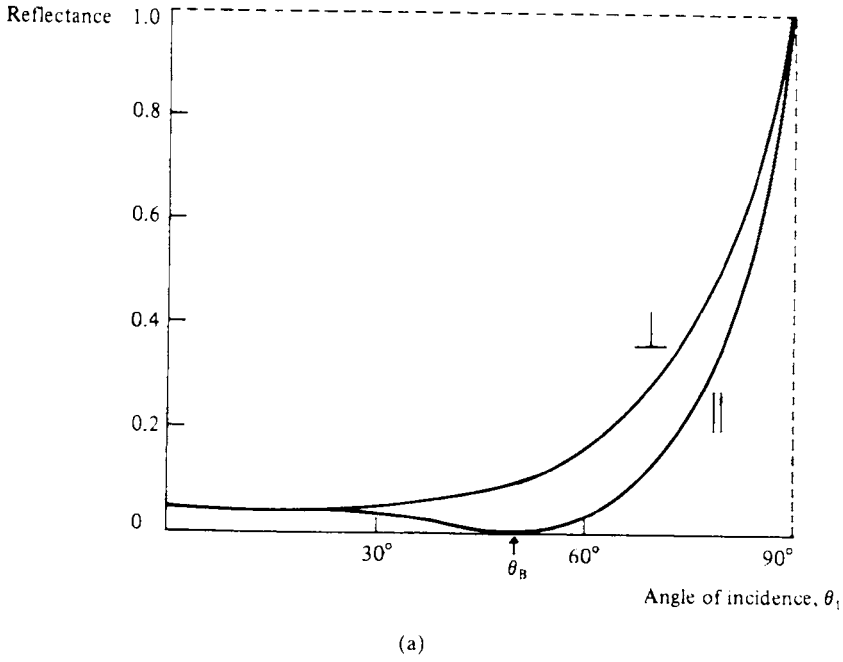


FIG. 1.6 Reflectance as a function of angle of incidence for light polarized parallel (||) and perpendicular (⊥) to the plane of incidence incident on the interface from (a) the less optically dense side and (b) the more optically dense side. The values of $\theta_B = 56.3^\circ$ (a) and 33.7° (b), and $\theta_c = 41.8^\circ$, are for an air ($n = 1$)/glass ($n = 1.5$) combination.

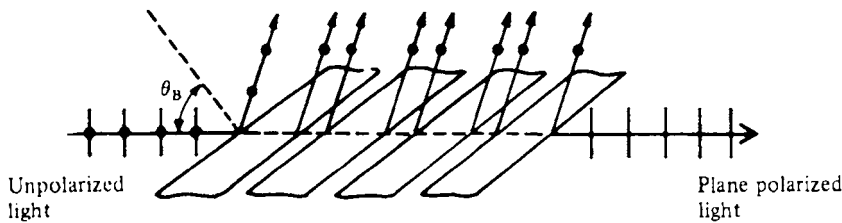


FIG. 1.7 Schematic illustration of the production of plane polarized light by reflection using a ‘pile of plates’. The light is incident on the plates at the Brewster angle so that with about six plates the emergent light is plane polarized parallel to the plane of incidence. For simplicity the effects of refraction at the surfaces of the plate are ignored.

EXAMPLE 1.1 Brewster angle

We may calculate the Brewster angle for glass given that its refractive index in air is 1.5.

From eq. (1.13) and assuming that the glass has an interface with air ($n = 1$) we have $\theta_B = \tan^{-1} 1.5 = 56.3^\circ$ (see Fig. 1.6).

In Fig. 1.5, and in eqs (1.12) and (1.13), we have tacitly assumed that $n_2 > n_1$, but for incidence at a boundary separating a more optically dense medium from a less optically dense medium the reverse is obviously the case. Hence the Brewster angle ‘in air’ is different from the Brewster angle ‘in the material’; for air ($n = 1$) to glass ($n = 1.5$) the two angles are $\tan^{-1} 1.5$ (or 56.3°) and $\tan^{-1} 0.67$ (or 33.7°). The reflectance of light polarized parallel or perpendicular to the plane of incidence varies with the angle of incidence in the more dense medium (i.e. $n_1 > n_2$) as shown in Fig. 1.6(b). We see here that for angles greater than a certain angle θ_c , the critical angle, there is no transmitted light, that is *total internal reflection* has taken place.

At $\theta = \theta_c$ the angle of refraction (in the less dense medium) is 90° , so from eq. (1.12) we can write

$$n_1 \sin \theta_c = n_2 \sin 90^\circ$$

or

$$\theta_c = \sin^{-1}(n_2/n_1) \tag{1.14}$$

Total internal reflection is discussed in more detail in relation to the propagation of light along optical waveguides in Chapter 8.

Light can also be polarized by selective absorption in various materials such as tourmaline, which occurs naturally, and ‘polaroid’. Polaroid consists of a plastic sheet of polyvinyl alcohol impregnated with iodine. Molecules of iodine polyvinyl alcohol are orientated into long chains by stretching the sheet. This material then transmits about 80% of the light polarized perpendicular to the chains of molecules but less than 1% of the light polarized parallel to these chains. Evidently the chains of molecules interact with this component and effectively absorb it. Polaroid is widely used for producing plane polarized light in optical

systems, but various polarizing prisms such as the Nicol prism and the Glan–Thompson prism are more efficient in terms of the degree of polarization (which is defined as the ratio of the irradiance of the polarized component to the total irradiance).

If the light incident on such a 'linear' polarizing device is already plane polarized, then the amount of light transmitted depends on the angle θ between the plane of polarization of the incident light and the plane of polarization of the light transmitted by the polarizer (often called the polarizing axis). We can easily show (see Problem 1.4) that if \mathcal{E}_0 is the amplitude of the incident light then the transmitted amplitude is $\mathcal{E}_0 \cos \theta$ and the irradiance of the transmitted light is given by³

$$I = \mathcal{E}_0^2 \cos^2 \theta = I_0 \cos^2 \theta \quad (1.15)$$

This relationship is known as *Malus's law*.

1.2.2 Principle of superposition

In discussing optical phenomena we are often confronted with the problem of finding the resultant produced when two or more waves act together at a point in space. It is found that the principle of superposition applies. This states that 'the resultant electric field at a given place and time due to the simultaneous action of two or more sinusoidal waves is the algebraic sum of the electric fields of the individual waves'. That is,

$$\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \dots$$

where $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are the electric fields of the individual waves at the specified time and place. The summation can be carried out in several different ways, the choice of method for a particular problem being a matter of mathematical convenience (ref. 1.2).

Let us consider the simple case of the superposition of two waves of the same frequency propagating in the same direction given by

$$\begin{aligned} \mathcal{E}_1 &= \mathcal{E}_{01} \sin(\omega t - kx + \phi_1) \\ \mathcal{E}_2 &= \mathcal{E}_{02} \sin(\omega t - kx + \phi_2) \end{aligned} \quad (1.16)$$

The resultant is

$$\begin{aligned} \mathcal{E} &= \mathcal{E}_1 + \mathcal{E}_2 \\ \mathcal{E} &= \mathcal{E}_{01} \sin(\omega t - kx + \phi_1) + \mathcal{E}_{02} \sin(\omega t - kx + \phi_2) \end{aligned}$$

which may be written as

$$\begin{aligned} \mathcal{E} &= (\mathcal{E}_{01} \cos \phi_1 + \mathcal{E}_{02} \cos \phi_2) \sin(\omega t - kx) \\ &\quad + (\mathcal{E}_{01} \sin \phi_1 + \mathcal{E}_{02} \sin \phi_2) \cos(\omega t - kx) \end{aligned}$$

This is identical with

$$\mathcal{E} = \mathcal{E}_0 \sin(\omega t - kx + \phi) \quad (1.17)$$

provided that

$$\mathcal{E}_0^2 = (\mathcal{E}_{01} \cos \phi_1 + \mathcal{E}_{02} \cos \phi_2)^2 + (\mathcal{E}_{01} \sin \phi_1 + \mathcal{E}_{02} \sin \phi_2)^2$$

or

$$\mathcal{E}_0^2 = \mathcal{E}_{01}^2 + \mathcal{E}_{02}^2 + 2\mathcal{E}_{01}\mathcal{E}_{02}\cos(\phi_2 - \phi_1) \quad (1.17a)$$

and

$$\tan \phi = \frac{(\mathcal{E}_{01} \sin \phi_1 + \mathcal{E}_{02} \sin \phi_2)}{(\mathcal{E}_{01} \cos \phi_1 + \mathcal{E}_{02} \cos \phi_2)} \quad (1.17b)$$

The form of eq. (1.17) implies that the resultant of adding two sinusoidal waves of the same frequency is itself a sinusoidal wave with the same frequency as the original waves given by eqs (1.16). The resultant is, of course, a solution of the wave equation (1.3). By repeated application of this process we can show that the resultant of any number of sinusoidal waves of the same frequency is a sinusoidal wave of that frequency. In this case the resultant is given by eq. (1.17), but now we have

$$\mathcal{E}_0^2 = \left(\sum_i \mathcal{E}_{0i} \cos \phi_i \right)^2 + \left(\sum_i \mathcal{E}_{0i} \sin \phi_i \right)^2$$

or

$$\mathcal{E}_0^2 = \sum_i \mathcal{E}_{0i}^2 + \sum_{\substack{i,j \\ i \neq j}} \mathcal{E}_{0i} \mathcal{E}_{0j} \cos(\phi_j - \phi_i) \quad (1.18)$$

and

$$\tan \phi = \frac{\sum_i \mathcal{E}_{0i} \sin \phi_i}{\sum_i \mathcal{E}_{0i} \cos \phi_i}$$

Now if the original light waves are from completely independent sources, including separated regions of an extended source, the phase difference $(\phi_j - \phi_i)$ in eq. (1.18) will vary in a random way such that the average value of $\cos(\phi_j - \phi_i)$ is zero. That is, for every possible positive value of phase difference there is a corresponding negative value. Therefore, the irradiance resulting from the superposition of the light from such independent sources is given by

$$\mathcal{E}_0^2 = \sum_i \mathcal{E}_{0i}^2 \quad (1.19)$$

The resultant irradiance is seen to be the sum of the irradiances which would be produced by the individual sources acting separately. In this case the sources are said to be *incoherent* and an area on which the light from such sources falls will be uniformly illuminated. If, on the other hand, the waves from the different sources maintain constant phase relationships they are said to be *coherent*. When an area is illuminated simultaneously by two or more coherent sources, the irradiance usually varies from point to point giving rise to *interference fringes*.

If we consider a surface illuminated by two coherent sources, we can see that as we move across the surface the relative distance to the sources changes and the phase difference also

changes. Hence the term $\cos(\phi_2 - \phi_1)$ varies periodically between plus one and minus one and the irradiance varies from a maximum value I_{\max} , where

$$\begin{aligned} I_{\max} &= \mathcal{E}_0^2 = \mathcal{E}_{01}^2 + \mathcal{E}_{02}^2 + 2\mathcal{E}_{01}\mathcal{E}_{02} \\ &= (\mathcal{E}_{01} + \mathcal{E}_{02})^2 \end{aligned} \quad (1.20)$$

when the waves are in phase and $\cos(\phi_2 - \phi_1) = +1$, to a minimum value I_{\min} , where

$$\begin{aligned} I_{\min} &= \mathcal{E}_{01}^2 + \mathcal{E}_{02}^2 - 2\mathcal{E}_{01}\mathcal{E}_{02} \\ &= (\mathcal{E}_{01} - \mathcal{E}_{02})^2 \end{aligned}$$

when the waves are exactly out of phase and $\cos(\phi_2 - \phi_1) = -1$. Thus a regular system of alternate bright and dark interference fringes is formed over the surface. Extending the argument to a large number of coherent sources we can see from eq. (1.20) that the resulting irradiance at any point where all the waves are in phase is

$$I = \left(\sum_i \mathcal{E}_{0i} \right)^2 \quad (1.21)$$

At a nearby point the resultant irradiance may be very small, or indeed zero, depending on the relative magnitudes of the terms \mathcal{E}_{0i} . The interaction of waves from coherent sources thus alters the spatial distribution of energy without altering the total amount. The topics of interference and coherence will be pursued further in the next section and section 6.5.1.3 respectively. The superposition of polarized waves is considered in section 3.1.

EXAMPLE 1.2 Maximum irradiance of several coherent and incoherent sources

Here we compare the maximum irradiance resulting from the superposition of equal contributions I from four (a) coherent sources and (b) incoherent sources.

(a) From eq. (1.21) we have $I_{\max} = n^2 I = 16I$, while (b) from eq. (1.19) we have $I_{\max} = nI = 4I$.

STANDING WAVES

Let us now consider the superposition of two waves of the same frequency travelling in opposite directions. The waves may be represented by the equations

$$\mathcal{E}_1 = \mathcal{E}_0 \sin(\omega t - kx)$$

and

$$\mathcal{E}_2 = \mathcal{E}_0 \sin(\omega t + kx)$$

where, for convenience, we have taken $\mathcal{E}_{01} = \mathcal{E}_{02} = \mathcal{E}_0$ and ϕ_1 and $\phi_2 = 0$. Then $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2$, and hence

$$\mathcal{E} = 2\mathcal{E}_0 \cos kx \sin \omega t \quad (1.22)$$

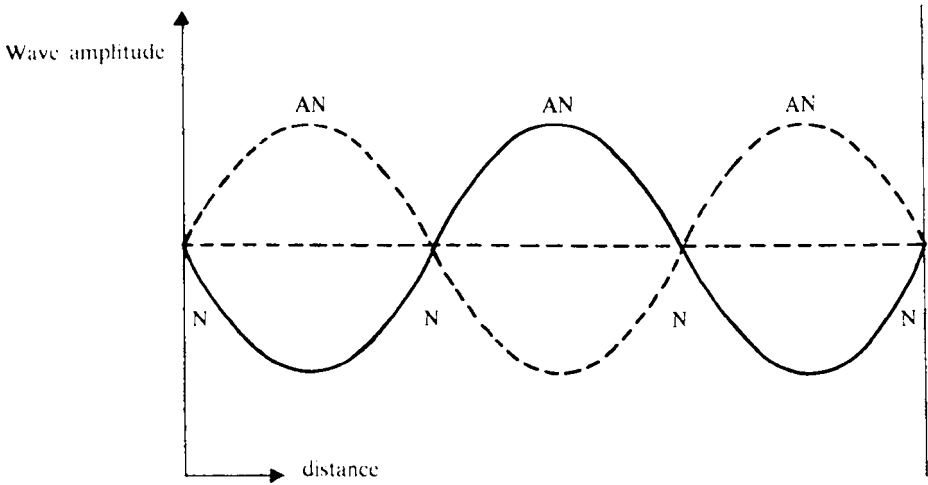


FIG. 1.8 Amplitude profile in a standing wave as a function of distance (N, nodes; AN, antinodes).

Equation (1.22) represents a standing wave and is illustrated in Fig. 1.8. In such a wave there is no net transfer of energy in either direction. The energy density in the medium, being proportional to the square of the amplitude of the motion (i.e. $4\epsilon_0^2 \cos^2 kx$), varies from a maximum at the antinodes to zero at the nodes. Standing waves of this type may be set up in optical cavities (or resonators), which consist of a pair of parallel, highly reflecting mirrors. Such a cavity forms an integral part of any laser.

1.2.3 Interference

The basic mathematical description of 'two-beam' interference is given by eq. (1.17a); that is,

$$I = \epsilon_{01}^2 = \epsilon_{01}^2 + \epsilon_{02}^2 + 2\epsilon_{01}\epsilon_{02} \cos(\phi_2 - \phi_1)$$

If $\epsilon_{01} = \epsilon_{02}$ then

$$I = 2\epsilon_{01}^2 [1 + \cos(\phi_2 - \phi_1)]$$

or

$$I = 4\epsilon_{01}^2 \cos^2\left(\frac{\phi_2 - \phi_1}{2}\right) \quad (1.23)$$

Equation (1.23) shows that the irradiance distribution of the fringes is given by a cosine-squared function. If the contributions from the coherent sources are equal the irradiance of the fringes varies from $4\epsilon_{01}^2$ to zero as $(\phi_2 - \phi_1)$ varies between 0 and π . If $\epsilon_{01} \neq \epsilon_{02}$, the resultant irradiance varies between $(\epsilon_{01} + \epsilon_{02})^2$ and $(\epsilon_{01} - \epsilon_{02})^2$.

To obtain the coherent wave trains required for the observation of interference before the

advent of lasers one had to ensure that:

1. The sets of wave trains were derived from the same small source of light and then brought together by different paths.
2. The differences in path were short enough to ensure at least partial coherence of the wave trains (i.e. the differences in path were less than the coherence length of the source – see section 6.5.1.3).

The basic ways of satisfying these requirements and demonstrating interference can be classified into two groups, namely 'division of wavefront' and 'division of amplitude'. The classic experiment of Young, in 1802, falls into the former group. In this experiment, which is illustrated in Fig. 1.9, monochromatic light is passed through a pinhole S so as to illuminate a screen containing two further identical pinholes or narrow slits placed close together. The presence of the single pinhole S provides the necessary mutual coherence between the light beams emerging from the slits S_1 and S_2 . The wavefronts from S intersect S_1 and S_2 simultaneously so that the light contributions emerging from S_1 and S_2 are derived from the same original wavefront and are therefore coherent. These contributions spread out from S_1 and S_2 (if these are long, thin slits) as 'cylindrical' wavefronts and interfere in the region beyond the screen. If a second screen is placed as shown then an interference pattern consisting of straight line fringes parallel to the slits is observed on it.

To find the irradiance at a given point P it is necessary to find the phase difference ϕ between the two sets of waves arriving at P from S_1 and S_2 . This in turn depends on the path difference $(D_2 - D_1)$ as in general

$$\text{phase difference} = \frac{2\pi}{\lambda_0} (\text{optical path difference})$$

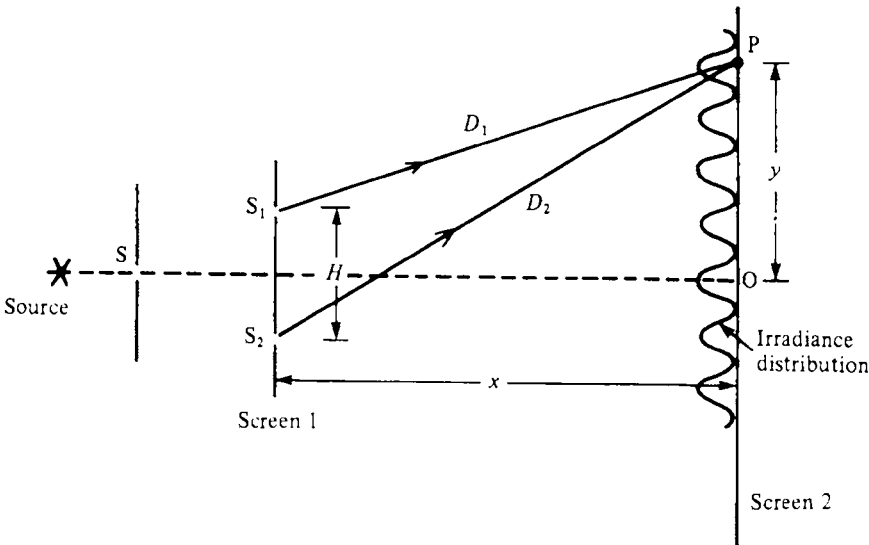


FIG. 1.9 Schematic layout and geometry for a Young's double slit interference experiment.

or using eq. (1.5),

$$\phi = \phi_2 - \phi_1 = \frac{2\pi}{\lambda} (D_2 - D_1)$$

where D_1 and D_2 are the distances from S_1 and S_2 to P respectively.

Bright fringes occur when the phase difference is zero or $\pm 2p\pi$, where p is an integer; that is, when

$$\frac{2\pi}{\lambda} (D_2 - D_1) = \pm 2p\pi$$

which is equivalent to $|D_2 - D_1| = p\lambda$. Therefore bright fringes occur if the path difference is an integral number of wavelengths. Similarly, dark fringes occur when $q = \pm(2p + 1)\pi$, or the path difference is an odd integral number of half-wavelengths. It is left as an exercise for the reader to show, with the parameters given in Fig. 1.9, that bright fringes occur at point P a distance y from O such that

$$y = \pm \frac{p\lambda x}{H} \quad (1.24)$$

provided both y and H are small compared with x . Here H is the slit separation and x is the distance from the screen containing the slits to the observing screen.

Equation (1.23) suggests that in such interference fringe patterns the irradiance of the fringe maxima will be equal. This is not the case, however, owing to diffraction effects, which are discussed in section 1.2.4.

Interference effects involving amplitude division can be observed in thin films or plates as illustrated in Fig. 1.10. In this case, interference occurs between the light reflected at A on the front surface of the plate and at B on the rear surface. If the plate has parallel faces then the two sets of waves from A and C are parallel and a lens must be used to bring them together to interfere. Using elementary geometry and Snell's law (eq. 1.12) the reader may show that the optical path difference $(AB + BC)n - AD$ is equal to $2nL \cos \theta_2$ where θ_2 is the angle of refraction and L is the plate thickness (see Problem 1.7). The phase difference is then $(2\pi/\lambda_0)(2nL \cos \theta_2)$ and therefore bright fringes occur when

$$\frac{4\pi nL \cos \theta_2}{\lambda_0} = 2p\pi$$

that is,

$$p\lambda_0 = 2nL \cos \theta_2 \quad (1.25)$$

where again λ_0 is the wavelength of the radiation in vacuum.

Likewise, dark fringes occur when

$$(2p + 1)\lambda_0/2 = 2nL \cos \theta_2 \quad (1.25a)$$

If the plate is optically denser than the surrounding medium, there is a phase change of π on reflection at the upper surface, thereby causing the above conditions to be interchanged.

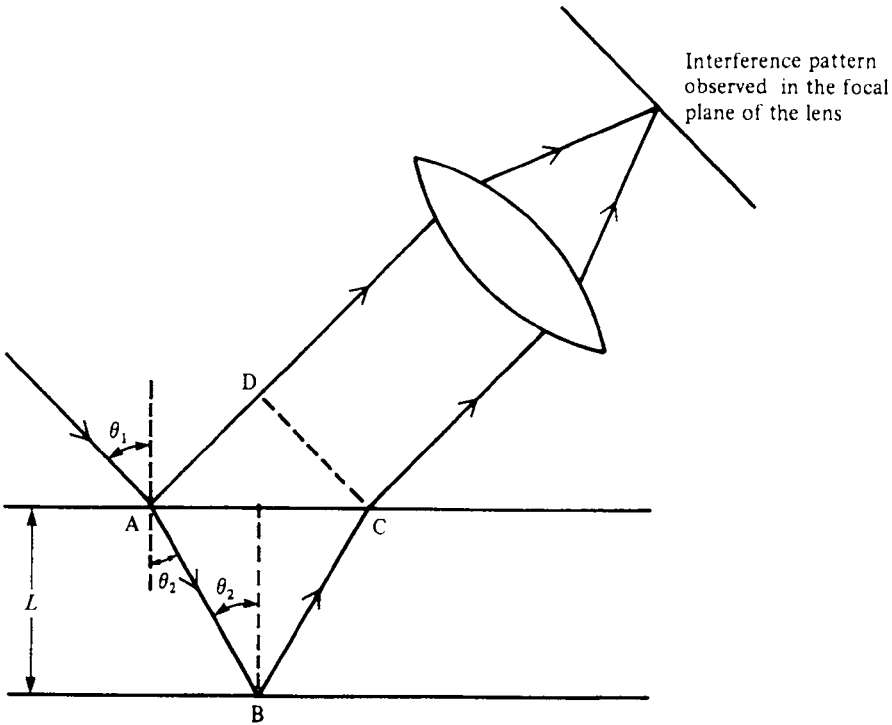


FIG. 1.10 Schematic diagram of interference effects in a thin film or plate of refractive index n .

For a given fringe p , λ_0 , L and n are constant and therefore θ_2 must be constant; the fringes are known as 'fringes of equal inclination'. If the angle of incidence is not too large and an extended monochromatic source is used then the fringes are seen as a set of concentric circular rings in the focal plane of the observing lens.

When the optical thickness of the plate is not constant and the optical system is such that θ_2 is almost constant, the fringes are contours of equal optical thickness nL . The situation may be illustrated by considering a small-angled wedge. If the wedge is uniform the fringes are approximately straight lines parallel to the apex of the wedge. Again it is left to the reader to show that the apex angle α is given by

$$\alpha \approx \tan \alpha = \frac{\lambda_0}{2S} \quad (1.26)$$

where S is the fringe spacing. The fringes are often close together and are conveniently viewed with a low power microscope.

MULTIPLE BEAM INTERFERENCE

If the reflectances of the surfaces of the plate shown in Fig. 1.10 were increased there would be many reflected beams to contribute to the interference pattern rather than just the two shown. In practice the resultant interference pattern is seen more clearly if the transmitted beams rather

than the reflected ones are used, as shown in Fig. 1.11. If the plate has parallel sides then the multiple beams are parallel and are brought together to interfere in the focal plane of the lens. The resultant of superposing these beams can be calculated quite easily as the phase difference from one beam to the next is constant. The phase difference is due to two additional traversals of the plate plus any phase changes which may occur on reflection at the surfaces of the plate. If the latter are ignored then the condition for the formation of fringe maxima is identical with the two-beam case given in eq. (1.25). The irradiance distribution is no longer the cosine-squared distribution of eq. (1.23), but rather it is given by (ref. 1.3)

$$I = \frac{I_0 T^2}{(1 - R)^2} \left(\frac{1}{1 + [4R/(1 - R)^2] \sin^2(\delta/2)} \right) \quad (1.27)$$

where I_0 is the irradiance of the incident beam, R and T are the reflectance and transmittance respectively of the plate surfaces and δ is the total phase change between successive beams. It should be noted that $T = 1 - R$, providing that there is no absorption in the plates.

If R is large, greater than about 0.8 say, then the fringe maxima are very sharp as shown in Fig. 1.12. Fringes of this type are formed in the Fabry–Perot interferometer. This instrument includes a pair of highly reflecting surfaces set accurately parallel to one another forming an optical (or Fabry–Perot) resonator as mentioned above and which is discussed further in section 5.5. The resolving power of the Fabry–Perot interferometer can be very high and it is widely used in the accurate measurement of the hyperfine structure of spectral lines.

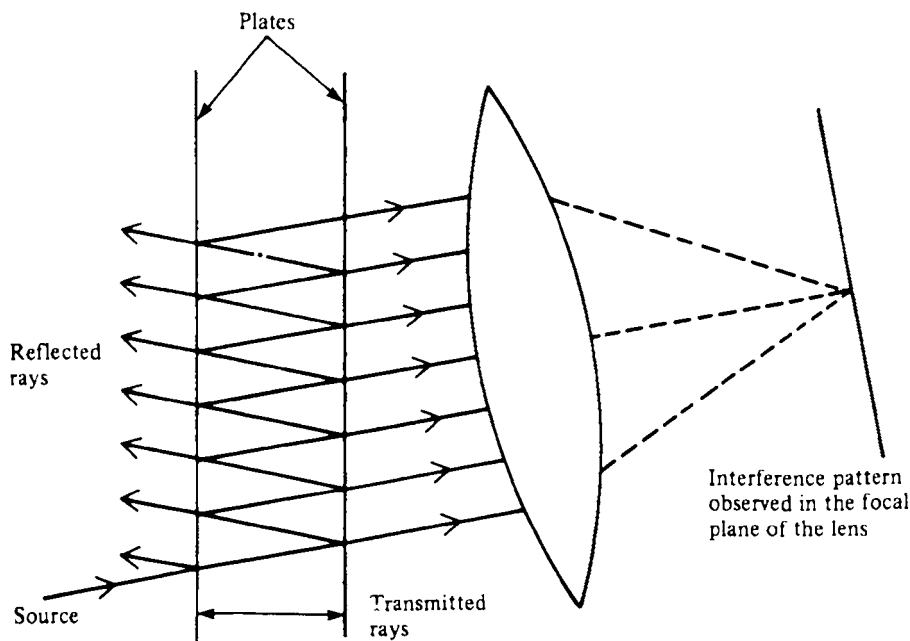


FIG. 1.11 Paths of light rays resulting from multiple reflections from two parallel plates (for simplicity the rear surfaces of the plates and refraction at the surfaces are not shown).

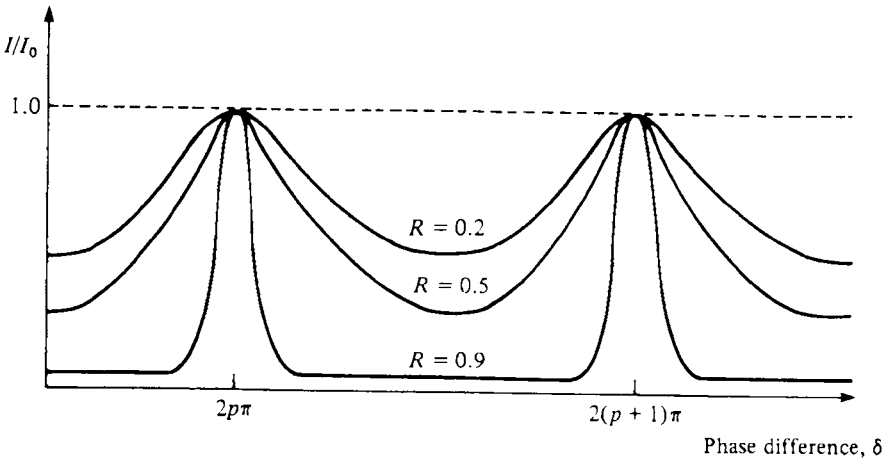


FIG. 1.12 Schematic diagram showing the variation in the irradiance distribution of multiple beam interference fringes as a function of mirror reflectance R .

Multiple beam interference effects can also be produced by division of wavefront. This is achieved by increasing the number of slits from two, the ultimate being reached in the diffraction grating where there is a very large number of equal slits. As before, the condition for interference maxima is that the contributions to the total irradiance emerging from successive slits should be in phase. It can be shown using elementary theory that, for normal incidence, maxima in the interference pattern produced by a diffraction grating occur when

$$p\lambda = (a + b)\sin \theta = d \sin \theta \quad (1.28)$$

where $d (= a + b)$ is the grating constant, with a and b the width and separation of the slits respectively and θ the angle of diffraction as shown in Fig. 1.13. For a more complete discussion of the theory of the diffraction grating the reader is referred to the texts given in ref. 1.1.

1.2.4 Diffraction

If an opaque object is placed between a source of light and a screen it is found that the shadow cast on the screen is not perfectly sharp. Some light is present in the dark zone of the geometrical shadow. Similarly light which emerges from a small aperture or narrow slit is observed to spread out. This failure of light to travel in straight lines is called *diffraction*; it is a natural consequence of the wave nature of light.

The essential features of diffraction can be explained by Huygens' principle, which states that:

The propagation of a light wave can be predicted by assuming that each point on the wavefront acts as a source of secondary wavelets which spread out in all directions. The envelope of these secondary wavelets after a small interval of time is the new wavefront.

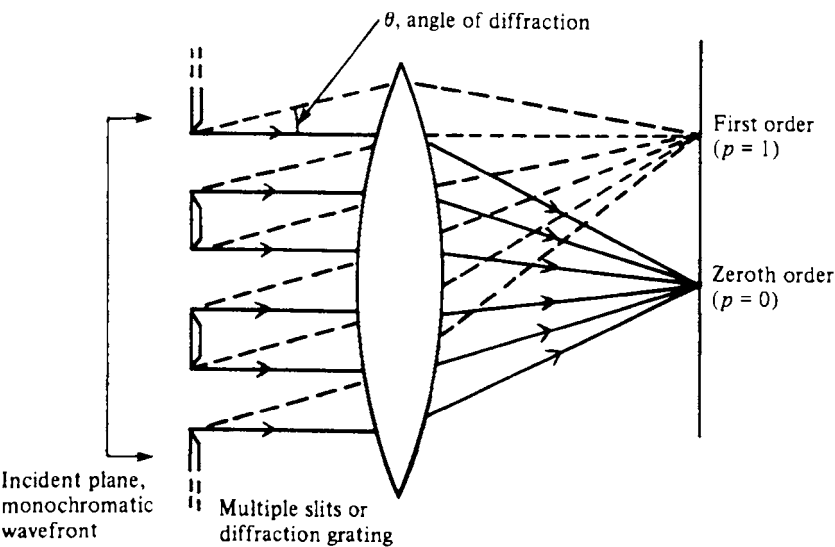


FIG. 1.13 Schematic illustration of the formation of interference maxima by a diffraction grating. If the incident light is not monochromatic then a first-, second- ... order interference fringe forms for each wavelength present at different values of θ so that each can therefore be distinguished.

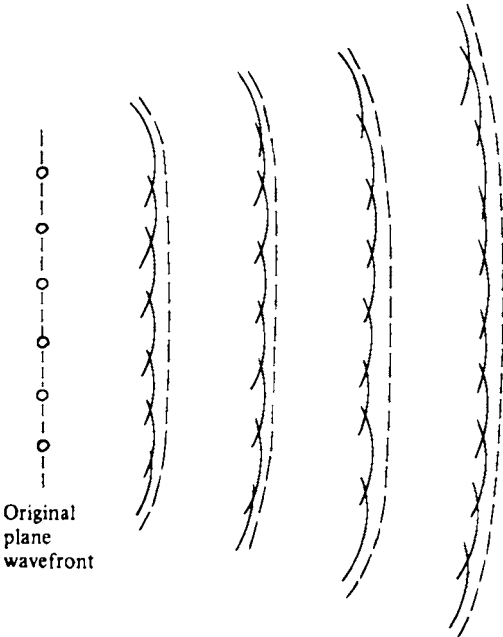


FIG. 1.14 Illustration of the application of Huygens' principle to the propagation of a plane wavefront (dashed lines). The small circles are typical 'point sources' on the wavefront which emit secondary wavelets. The wavefront at a later time is the resultant of the secondary wavelets. The diagram shows (in an exaggerated way) that an initially plane wavefront will diverge as it propagates.

In effect the spherical wavelets emitted from the point sources are summed using the principle of superposition; the result of the summation is equivalent to the wavefront at a later time. The propagation of a plane wave according to Huygens' principle is illustrated in Fig. 1.14. We see that the wavefronts develop some curvature at the edges due to the radiation from the end points being directed away from the axis. Succeeding wavefronts become more and more curved so that the beam diverges.

A quantitative description of diffraction can be obtained by setting Huygens' principle in a mathematical form known as the *Fresnel–Kirchhoff formula* (ref. 1.4).

In the detailed treatment of diffraction it is customary to distinguish between two general cases known as *Fraunhofer* and *Fresnel* diffraction. Qualitatively, Fraunhofer diffraction occurs when the incident and diffracted waves are effectively plane, while in Fresnel diffraction the curvature of the wavefront is significant. Clearly there is not a sharp distinction between the two cases.

Of particular importance is the Fraunhofer diffraction produced by a narrow slit; the arrangement for observing this is shown in Fig. 1.15(a). Over 80% of the light passing through the aperture falls within the central maximum of the pattern, on either side of which there is a series of low irradiance secondary maxima. These maxima are separated by minima which occur at angles of diffraction θ given by

$$\sin \theta = \frac{p\lambda}{D} \quad (1.29)$$

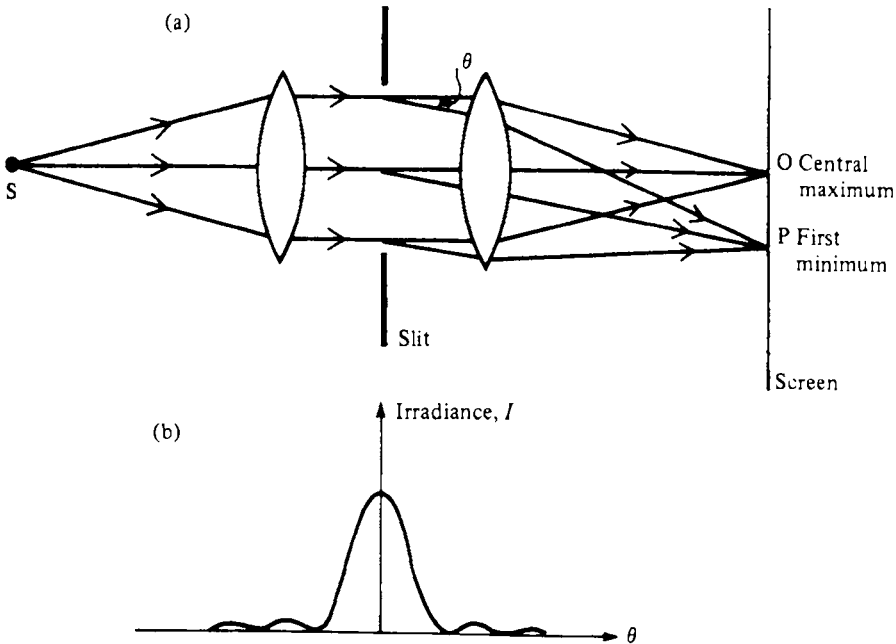


FIG. 1.15 Schematic arrangement for observing the Fraunhofer diffraction produced by a single slit or aperture (a) and the resultant irradiance distribution as a function of the angle of diffraction θ (b).

where D is the width of the slit. A rather more important case is that of the circular aperture which produces a pattern consisting of a central bright area surrounded by concentric dark and bright rings. About 84% of the light is concentrated within the central spot, which is called the *Airy disk*. A measure of the amount of diffraction is given by the angle θ at which the first dark ring occurs. It may be shown, using the Fresnel – Kirchhoff formula, that θ is given by

$$\sin \theta = 1.22 \frac{\lambda}{D} \approx \frac{\lambda}{D} \quad (1.30)$$

where D is the diameter of the aperture. In many applications the output from a laser or other source is focused, expanded, or otherwise passed through lenses, prisms, stops and the like. In each case the edges of the optical component may serve as an aperture limiting the extent of the beam and thereby introducing additional divergence due to diffraction.

In passing, it is noteworthy that diffraction sets a theoretical limit to the resolving power of optical instruments and it is also responsible for the occurrence of ‘missing orders’ in the interference patterns produced by double and multiple split arrangements. The irradiance distribution of such patterns has an envelope which is the single slit diffraction pattern shown in Fig. 1.15(b). Thus at certain angles θ where an interference maximum is expected none appears because a diffraction minimum occurs at the same angle.

EXAMPLE 1.3 Resolving power: the Rayleigh criterion

We may estimate (a) the minimum separation of two point sources that can just be resolved by a telescope with an objective lens of 0.1 m diameter which is 500 m from the sources and (b) the minimum wavelength difference which may be resolved by a diffraction grating, which is 40 mm wide and has 600 lines mm^{-1} , in the first order. (Assume $\lambda = 550 \text{ nm}$ in both cases.)

We define resolving power in terms of the Rayleigh criterion, which states that two objects (or wavelengths) are just resolved if their diffraction patterns when viewed through the optical system are such that the principal maximum of one falls in the first minimum of the other.

In (a) this criterion therefore requires that the sources have an angular separation of $\theta \approx \lambda/D$ (see eq. 1.29), that is $\theta \approx \lambda/D$. Now θ is also given by $\theta = S_{\min}/500$, where S_{\min} is the minimum separation of the sources. Hence $\lambda/D = S_{\min}/500$, whence

$$S_{\min} = \frac{500 \times 550 \times 10^{-9}}{0.1} = 2.75 \text{ mm}$$

For (b) the chromatic resolving power $\lambda/\delta\lambda$ of a grating is given by $\lambda/\delta\lambda = pN$, where N is the number of lines used. Therefore, in the first order $p = 1$, $\lambda/\delta\lambda = 1 \times 40 \times 600$; that is, $\delta\lambda = 0.023 \text{ nm}$.

1.3

Optical components

Various optical components such as mirrors and lenses are widely used in optoelectronics. For example, as we shall see in Chapter 5 the operation of lasers is crucially dependent on an optical system comprising two plane or spherical mirrors, whilst lenses are often used to collect, collimate and focus the light from diverging sources such as light-emitting diodes (section 4.6). Mirrors, lenses and prisms are also used to manipulate and steer beams of radiation within a variety of optical/optoelectronic systems.

In this section we present the various formulae dealing with the operation of such components, the appropriate derivations of which may be found in most elementary optics textbooks (see e.g. ref. 1.1).

1.3.1 The spherical mirror

Figure 1.16 shows a ray from a source S on the axis of a spherical mirror at a distance s from its vertex V . The ray is redirected by the mirror to pass through the axis at the point T , a distance s' from V . Provided the rays are close to the axis (i.e. provided they are 'paraxial' rays, ref. 1.5), then the distances s and s' are related by the equation

$$\frac{1}{s} + \frac{1}{s'} = \frac{2}{r} \quad (1.31)$$

where r is the radius of curvature of the mirror.

If the incident rays are parallel to the axis, that is $s = \infty$, then the image is formed at the

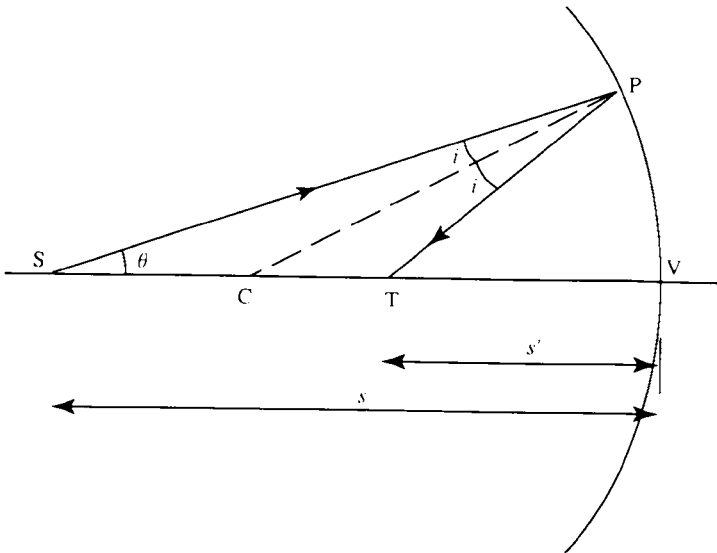


FIG. 1.16 Diagram showing the reflection of light from an object S on the axis of a spherical mirror to form an image at T .

focal point of the mirror F, which is a distance f from V. Thus from eq. (1.31) putting $f=2r$ we have

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{1.32}$$

Similarly if the object is placed at F, the image is formed at infinity. Likewise if a small source of light is located at F, a parallel beam will be generated.

In applying equations such as eq. (1.32) it is necessary to adopt a sign convention for the distances involved. Of the several available a convenient one is to adopt the signs used in coordinate geometry. Consequently distances measured from the vertex are positive if measured to the right and negative to the left, and similarly distances measured upwards and downwards from the axis are positive and negative respectively. Thus in eq. (1.32) s , s' and f are all negative so the form of the equation remains unchanged.

The transverse magnification, m , of the mirror can be shown to be given by

$$m = -s'/s \tag{1.33}$$

the minus sign indicating that the image is inverted (Fig. 1.17).

1.3.2 The thin spherical lens

Figure 1.18 shows the path of a ray which originates from a point at S in a medium of refractive index n_1 as it passes through a simple spherical lens. The latter has two spherical faces with radii r_1 and r_2 which are separated by a distance d along the axis and it is made from a material with a refractive index of n_2 . The ray is diffracted as it passes through the two surfaces of the lens and subsequently crosses the axis at a point T. Thus an object placed at S will give rise to an image at T. If the distances from the lens surfaces to the points S and T are s and s' and the thickness of the lens is negligible compared with s and s' , then we have that

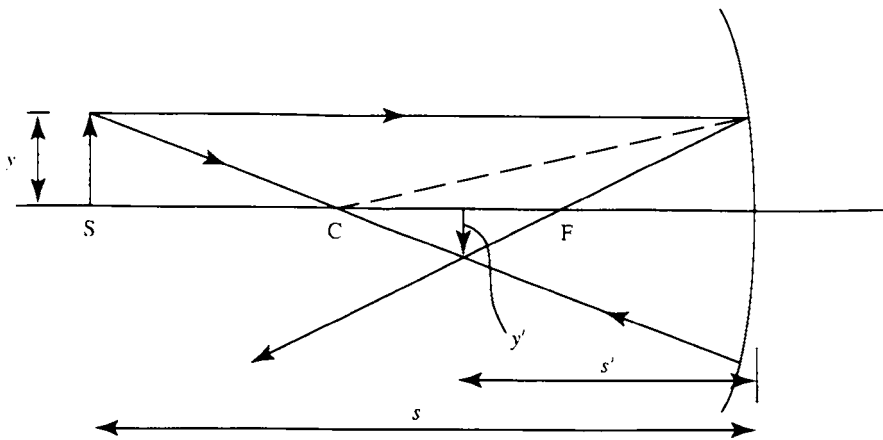


FIG. 1.17 Diagram showing the formation of an image of an off-axis object, which can be used to determine the magnification y'/y .

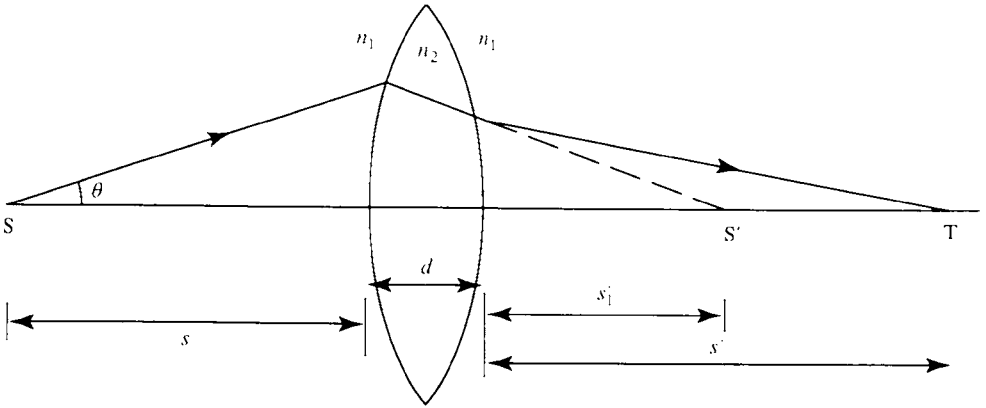


FIG. 1.18 Schematic diagram showing the behaviour of a double convex lens.

$$\frac{n_1}{s'} - \frac{n_1}{s} = (n_2 - n_1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (1.34)$$

In most cases of practical importance the object and image are in air so that we may put $n_1 = 1$. Writing $n_2 = n$ and noting that when $s' = \infty$, $s = f$, where f is the focal length (similarly if $s = \infty$, $s' = f'$), we have in general

$$\frac{1}{s'} - \frac{1}{s} = \frac{1}{f'} - \frac{1}{f} = -\frac{1}{f} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (1.35)$$

where f and f' , which are numerically equal, are often referred to as the first and second focal lengths respectively. Frequently, however, we simply refer to the focal length of a lens meaning the second focal length; the focal lengths of converging lenses such as that shown in Fig. 1.19 are then positive, whilst those of diverging lenses are negative.

By considering the formation of the image of an object which extends beyond the axis as shown in Fig. 1.19, the lateral magnification of the lens, m , may be shown to be given by an equation which is identical to eq. (1.33):

$$m = -s'/s$$

As the object becomes larger, and/or as the diameter or aperture of the lens increases, some of the rays of light from the object to the lens will increasingly violate the 'paraxial' condition leading to a reduction in the quality of the image. Such effects are referred to as *monochromatic aberration*, the most familiar form of which is *spherical aberration* (ref. 1.6).

Equation (1.35) shows that the focal length is dependent on the refractive index of the lens material, which in turn is a function of wavelength. Thus images formed by simple lenses often have a coloured boundary due to this *chromatic aberration*.

It is worth noting that optical components inevitably have a finite aperture and can therefore only collect a fraction of the wavefronts from the object. Consequently there will always be an apparent deviation from rectilinear propagation when light passes through a lens, or

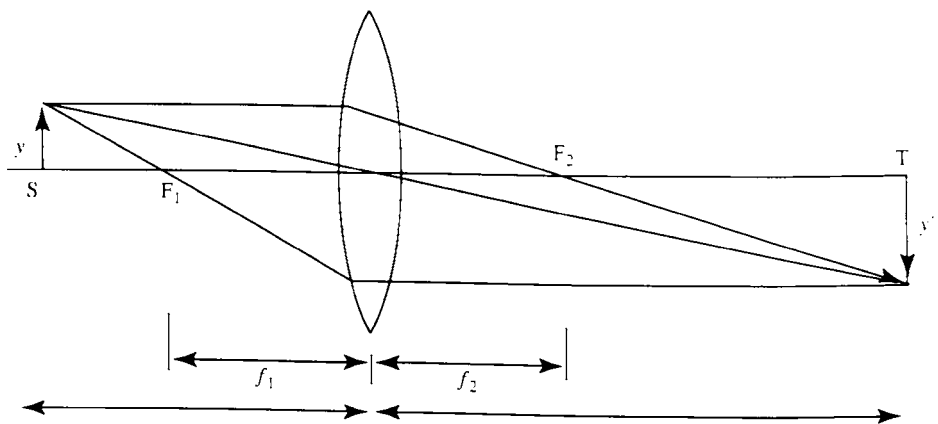


FIG. 1.19 Diagram showing the formation of an image at T, by a biconvex lens, of an object at S, which extends away from the optic axis. The diagram enables the lateral magnification $m = y'/y$ to be evaluated in terms of the object and image distances.

other aperture, so that the waves will be diffracted; the perfection of the image is thus said to be *diffraction limited*.

1.3.3 Other lenses

In practice simple thin lenses seldom completely satisfy the requirements placed upon them, and there is a wide range of *compound* lenses available, which comprise two or more lenses either in contact or separated by small distances. Such lens ‘systems’ may correct for the aberrations mentioned above, or yield very short focal lengths and/or very high magnifications. Typical of these is the microscope objective lens which is often used in the laboratory to focus light into very small apertures. Other lenses used for this purpose include the gradient index or GRIN lens and sphere. Small spheres, some 2–5 mm in diameter, are easier to manufacture than comparable lenses and are easier to align in experimental applications. The focal length of a sphere lens lies just outside of the sphere, and Fig. 1.20 shows the use of two such spheres for the collection, collimation and refocusing of the light from a small source.

GRIN lenses use materials in which the refractive index varies in some controlled way so that, by combining the refraction which occurs at the surface of the lens with that which occurs continuously as a function of the varying refractive index, an image can be produced. A typical GRIN lens is a cylindrical rod in which there is a radial variation of the refractive index, which is given by (ref. 1.7)

$$n(r) = n_0 \left(1 - \frac{A}{2} r^2 \right) \tag{1.36}$$

where n_0 is the refractive index on the lens axis, and A is a positive constant. As a result of the refractive index variation, rays of light incident on the end of the cylindrical lens follow

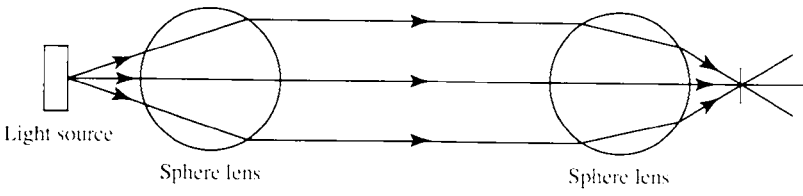


FIG. 1.20 Collection, collimation and refocusing of light using two sphere lenses.

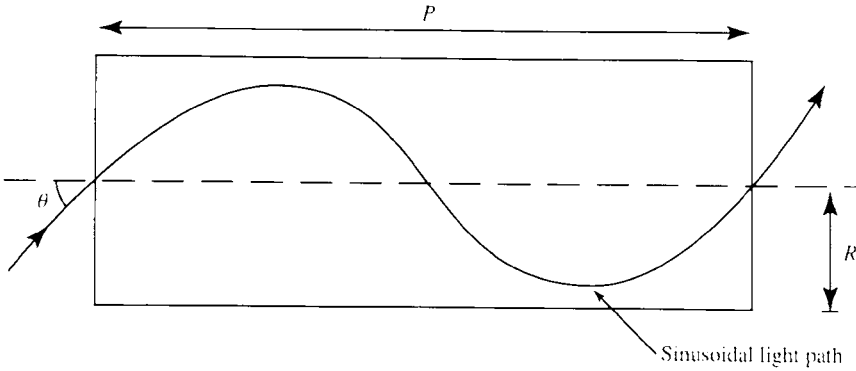


FIG. 1.21 Schematic diagram showing the path of a ray of light through a GRIN lens.

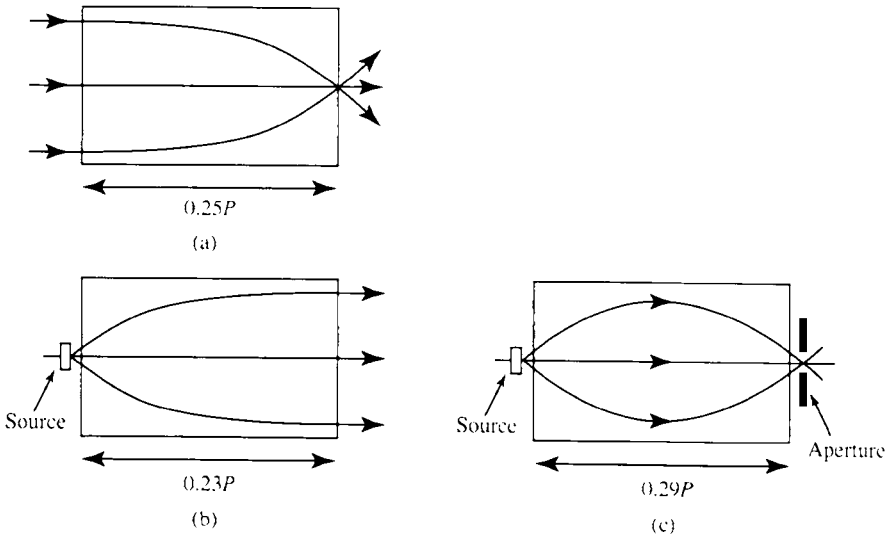


FIG. 1.22 Operation of typical GRIN lenses: (a) the quarter-pitch lens; (b) the $0.23P$ lens, used to collect the light diverging from a small source and collimate it into a parallel beam; (c) the $0.29P$ lens used, for example, to couple the output from a source into a small aperture.

sinusoidal paths as illustrated in Fig. 1.21, with a period or pitch P of the path given by $P = 2\pi/\sqrt{A}$, while the maximum acceptance angle θ of the lens is given by

$$\sin \theta = n_i R \sqrt{A} \quad (1.37)$$

where R is the lens radius. The characteristics of a GRIN lens in terms of the relationship between the image and object distances are determined by the length of the lens relative to the pitch. The quarter-pitch ($0.25P$) lens produces a perfectly collimated output beam when the light emanates from a point source on the opposite face of the lens. Conversely the lens focuses a collimated beam to a point at the centre of the opposite face (Fig. 1.22a), that is the focal point of the quarter-pitch GRIN lens is coincident with the lens faces. On the other hand Figs 1.22(b) and (c) show GRIN lenses which are either just shorter ($0.23P$) or just longer ($0.29P$) than a quarter pitch. In these cases the focal points lie just outside the lens and are therefore easily accessible in terms of locating a source of light or an image. Figure 1.22(c) illustrates how light from a small source can be coupled into a small aperture, when both are located close to the ends of the lens. The dimensions of such lenses are typically 2.0 mm in diameter by 6 mm in length, while the acceptance angle is typically 20° – 28° .

1.4

Light sources – blackbody radiation

The sources discussed in this section are the so-called classical or *thermal sources*. These are so named because they radiate electromagnetic energy in direct relation to their temperature. Thermal sources can be divided into two classes, namely blackbody radiators and line sources. The former are opaque bodies or hot, dense gases which radiate at virtually all wavelengths. Line sources, on the other hand, radiate at discrete wavelengths.

1.4.1 Blackbody sources

The radiation from opaque objects and dense gases was widely studied in the late nineteenth century resulting in the formulation of the following empirical laws. First, it was found that the optical power radiated from a body was proportional to the fourth power of the absolute temperature, T , and could be written as

$$W = \epsilon \sigma A T^4 \quad (1.38)$$

where A is the area of the body and σ the Stefan–Boltzmann constant. The parameter ϵ is called the emissivity; its value lies between zero and unity depending on the nature of the surface of the emitting body. The ‘ideal’ emitter has an emissivity of unity and is known as a *blackbody*. The name arises because it can be shown that a body whose emissivity is unity will completely absorb any radiation falling upon it; hence it will appear ‘black’. An approach to an ideal blackbody emitter can be made by piercing a small hole in an otherwise closed cavity, and then, if the cavity is maintained at a uniform, constant temperature, the radiation leaving the hole is essentially that of a blackbody. Most hot surfaces only approach the ideal and consequently have emissivity values of less than unity.

EXAMPLE 1.4 Radiated power

We may estimate the total power radiated from a source of area 10^{-5} m^2 at a temperature of 2000 K, given that the emissivity of the surface is 0.7.

From eq. (1.38) we have

$$W = 0.7 \times 5.67 \times 10^{-8} \times 10^{-5} \times (2000)^4 = 6.35 \text{ W}$$

Secondly it was noted that the spectral distribution of the energy emitted at a given temperature has a definite maximum and that this maximum shifts to shorter wavelengths as the temperature increases, as illustrated in Fig. 1.23. This shift is given by *Wien's displacement law*, which can be expressed as

$$\lambda_m T = \text{constant} \quad (1.39)$$

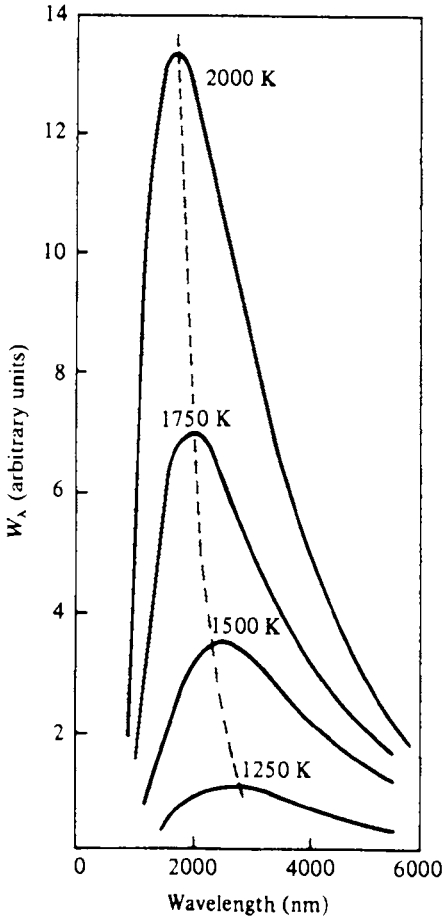


FIG. 1.23 Distribution of energy in the spectrum of a blackbody radiator at various temperatures.

where λ_m is the wavelength at which the radiated power is a maximum for a given temperature T .

The spectral distribution of blackbody radiation can be described in terms of the *spectral radiant emittance* function, $W(\lambda, T)$, where the power radiated from a blackbody per unit area between the wavelengths λ and $\lambda + d\lambda$ is given by $W(\lambda, T) d\lambda$. Planck derived the following functional form for $W(\lambda, T)$:

$$W(\lambda, T) = \frac{2\pi hc^3}{\lambda^5} \left(\frac{1}{\exp(hc/\lambda kT) - 1} \right) \quad (1.40)$$

Figure 1.23 shows the variation in $W(\lambda, T)$ with wavelength at a number of different temperatures.

Sometimes it is useful to consider the radiation distribution as a function of frequency rather than wavelength, when we have

$$W(\nu, T) = \frac{2\pi h \nu^3}{c^2} \left(\frac{1}{\exp(h\nu/kT) - 1} \right) \quad (1.41)$$

We have seen that a close approximation to a blackbody can be constructed by piercing a hole in an otherwise closed cavity: the smaller the hole, the nearer the emitted radiation approaches that of a blackbody. This naturally leads to a consideration of the radiation within a totally sealed container. If $\rho(\nu, T)d\nu$ denotes the energy density of such radiation between the frequencies of ν and $\nu + d\nu$ where the cavity is at a temperature T , then it may be shown (see Problem 1.11) that

$$\rho(\nu, T) = \frac{4}{c} W(\nu, T)$$

Substituting for $W(\nu, T)$ from eq. (1.41) gives another form of Planck's equation which will be found useful in Chapter 5, that is

$$\rho(\nu, T) = \frac{8\pi h \nu^3}{c^3} \left(\frac{1}{\exp(h\nu/kT) - 1} \right) \quad (1.42)$$

In the derivation of this equation, Planck considered the possible standing wave patterns or modes which can exist within a cavity and assumed that the energy associated with each mode was quantized; that is, the energy could only exist in integral multiples of some lowest amount or quantum. Thus, according to Planck, matter could only emit discrete quantities of radiation which were called photons. The success of this assumption, as we mentioned in the introduction to this chapter, provided the foundation for the development of modern quantum theory.

1.4.2 Line sources

In the case of excited gases in which there is little interaction between the individual atoms, ions or molecules, the electromagnetic radiation is emitted at well-defined wavelengths. This can be understood quite easily on the basis of the simple Bohr model of the atom in which

it is considered that the atom consists of a positive nucleus of charge Ze (Z being the atomic number of the atom) with electrons of mass m and charge e in certain 'allowed' bound orbits around it. Each of these orbits corresponds to a well-defined energy level. The energy is given by

$$E_n = \frac{-mZ^2e^4}{8n^2h^3\epsilon_0^2} \quad (1.43)$$

where n is an integer known as the *principal quantum number* (see e.g. ref. 2.1b). The outermost electron may be excited from its normal or ground state orbit to higher energy orbits which are normally unoccupied. When an electron undergoes a transition from one of these excited orbits (or energy levels) to a lower orbit it emits a quantum of radiation. The energy of the quantum is just the difference ΔE between the energies of the initial and final orbits. Thus the quantum energy is

$$h\nu = \frac{hc}{\lambda} = \Delta E \quad (1.44)$$

from which, using eq. (1.43), we see that

$$\nu = \frac{\Delta E}{h} = \frac{mc^4Z^2}{8h^3\epsilon_0^2} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right)$$

n_f and n_i being the values of the principal quantum number corresponding to the final and initial orbits (or energy levels) involved in the transition.

The spectral lines emitted in this way can have a very narrow frequency spread; that is, they are very nearly monochromatic. In practice, however, there are a number of causes of spectral broadening which increase the spread of the wavelength (or frequency) associated with the emitted photons (see section 5.7).

As long as the atoms are in thermal equilibrium with their surroundings, the energy radiated by an intense line radiator can never exceed that of a blackbody at the same temperature as the line source. This is true despite the very different wavelength distributions of the emitted energy of the two sources, even when comparing the energies they emit per unit wavelength range (or their spectral radiant emittances). This rule is broken in the case of lasers, where, as we shall see in Chapter 5, the atoms are *not* in thermal equilibrium.

EXAMPLE 1.5 Ionization energy of the hydrogen atom

Here we calculate the ionization energy of the hydrogen atom ($Z=1$) given the physical constants in Appendix 6. The ionization energy is the energy required to excite the electron from the ground state ($n=1$) to infinity ($n=\infty$).

Thus from eq. (1.43) we have

$$\begin{aligned} E_{\text{ion}} &= \frac{9.1 \times 10^{-31} \times (1.6 \times 10^{-19})^4}{8 \times (6.6 \times 10^{-34} \times 8.85 \times 10^{-12})^2} \\ &= 2.176 \times 10^{-18} \text{ J or } 13.6 \text{ eV} \end{aligned}$$

In concluding this section we return briefly to the photoelectric effect, which dramatically illustrates the photon or particle nature of light. Einstein explained the emission of electrons from metal surfaces exposed to light as being due to a transfer of energy from a single photon to a single electron. It was found that (a) the energy of the emitted electrons depends not on the irradiance of the incident light but rather on its frequency, (b) for light of a given frequency the photoelectrons have a maximum kinetic energy $E_{\max} (= \frac{1}{2} m v_{\max}^2)$ and (c) for a given metal there is a minimum, or threshold, frequency ν_0 for the light which will cause electrons to be emitted. These observations are summarized in Einstein's photoelectric equation

$$E_{\max} = h\nu - e\phi \quad (1.45)$$

where e is the electron charge and ϕ is a constant for a particular metal known as the *work function*. The quantity $e\phi$ represents the energy required to free an electron from the surface (section 2.6). The difference between the incident photon energy $h\nu$ and $e\phi$ then appears as the kinetic energy of the emitted electron.

EXAMPLE 1.6 Work function of metals

Given that the lowest frequency of light which will eject electrons from a tungsten surface is 1.1×10^{15} Hz, we may calculate its work function.

From eq. (1.45) we have $h\nu_0 = e\phi$, where ν_0 is the *threshold frequency*, and we assume that the ejected electrons have zero kinetic energy.

Hence $\phi = 4.5$ eV.

1.5 Units of light

The measurement of the energy of electromagnetic radiation when all wavelengths are treated equally is known as *radiometry*. The measurement of those aspects of radiation which affect vision is referred to as *photometry*. The link between these is the standard luminosity curve shown in Fig. 1.24, which shows the spectral response V_λ of the average eye to light of different wavelengths. The value of V_λ (often called the relative luminous efficiency) is taken as unity at $\lambda = 555$ nm where the eye has its maximum sensitivity. The value of V_λ falls to near zero at the extremes of the visible spectrum; that is, at about 400 nm and 700 nm. For normal photopic vision (when the eye is adapted for high levels of stimulus) at the peak sensitivity of the eye (555 nm) 1 watt of radiant energy equals 680 lumens by definition. The watt is a radiometric unit whereas the lumen is a photometric unit. At any other wavelength this conversion is scaled by the value of the relative luminous efficiency at that wavelength.

To obtain a measure of the relative brightness of various sources we must use photometric units. In laser technology and safety, however, radiometric units are widely used. We shall therefore briefly describe the most important radiometric units and give their photometric equivalents in Table 1.2.

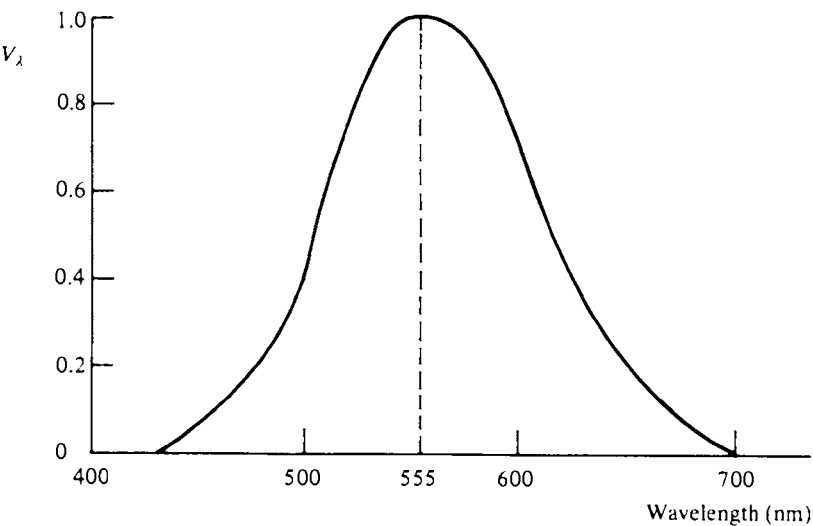


FIG. 1.24 Relative luminous efficiency curve for normal photopic vision (i.e. when the eye is adapted for high levels of stimulus).

TABLE 1.2 Radiometric and photometric units

Symbol (SI units)	Radiometric term and units	Photometric term and units	Definition
Q	Radiant energy (J)	Luminous energy (talbot)	
Φ	Radiant power or flux (W)	Luminous power or flux (lm)	
$I(E)$	Irradiance (W m^{-2})	Illuminance (lm m^{-2})	Total power falling on unit area
$J(I)$	Radiant intensity (W sr^{-1})	Luminous intensity (lm sr^{-1})	Power radiated by a point source into unit solid angle
L	Radiance ($\text{W m}^{-2} \text{sr}^{-1}$)	Luminance or brightness ($\text{lm m}^{-2} \text{sr}^{-1}$)	Radiant/luminous intensity per unit projected area in a given direction
W	Radiant emittance (W m^{-2})		Total power radiated in all directions from unit area

Note: Any of these quantities can be expressed per unit frequency or wavelength. In these cases the word ‘spectral’ is added as a prefix to that term and a ν or λ is added as a subscript to the symbol. Thus, for example, the spectral radiance L_λ is the radiance divided by the bandwidth in wavelength units (μm , nm , etc.). The spectral radiance may then have units of $\text{W m}^{-2} \text{sr}^{-1} \text{nm}^{-1}$ or, if the bandwidth is given in frequency units, the spectral radiance would have units $\text{W m}^{-2} \text{sr}^{-1} \text{Hz}^{-1}$.

Of course, light is a form of energy, and radiant energy and power are measured in joules and watts respectively. Radiant power is sometimes referred to as *flux* Φ ; the flux per unit area delivered to a surface is the *irradiance* I . It should be noted at this point that many texts on optics and lasers call the power per unit area the *intensity*. As we mentioned in footnote 1, however, this practice should not be encouraged and accordingly in this text we shall use

irradiance. The recommended symbol for irradiance is E (or E_e) but to avoid confusion with E used for energy we shall use the symbol I .

The energy emitted from a point source is described in terms of the *radiant intensity* \mathcal{I}_e , which is the radiant flux emitted by a *point* source within a unit solid angle in a given direction. \mathcal{I}_e is therefore measured in watts per steradian (W sr^{-1}).

For small *plane* sources, the equivalent term to radiant intensity is the *radiance* L_e , which is the power radiated into unit solid angle, in a given direction per unit projected area of source; that is, L_e has units $\text{W m}^{-2} \text{sr}^{-1}$. Several laser texts refer to this quantity as the *brightness*, which is possibly the most frequently used and also the most confusing term in this subject. The photometric quantity corresponding to irradiance is *illuminance*, defined as the flux delivered to a surface per unit area. The visual effect of this energy is characterized by the *luminance* or *brightness*. Brightness is also often used to describe psychological perception. The safest procedure would appear to be to take careful note of the *units* of the terms used in a given discussion.

The more commonly used radiometric quantities and their photometric (or luminous) equivalents are shown in Table 1.2. To reduce the profusion of nomenclature which existed in the recent past, the same symbols are used for corresponding radiometric and photometric quantities. When it is necessary to distinguish between radiant and luminous quantities the subscripts *e* (for energy) and *v* (for visual) are used.

The photometric quantity luminous intensity \mathcal{I}_v is one of the seven quantities chosen as dimensionally independent *base quantities* in the SI system of units. The units of luminous intensity, namely lumens per steradian (lm sr^{-1}), are often referred to as candela (*cd*). The luminance, therefore, has units of cd m^{-2} or, more commonly in practice, cd cm^{-2} despite the centimetre not being a recognized SI unit. Quite often Imperial units are used and the situation is further complicated depending on whether the emitting or reflecting surface is a Lambertian surface or a uniform diffuser – see section 4.8 and ref. 1.8.

NOTES

1. In the past the word *intensity* has been used for the flow of energy per unit time per unit area. However, by international agreement that term is being replaced by *irradiance*.
2. In the literature on optics one meets the terms reflectivity, reflection coefficient and reflectance. Unfortunately, these terms are often defined in different ways. For example, reflectance is often used for the ratio of the reflected energy per unit area (or flux) to the incident flux and also for the ratio of the reflected to the incident amplitude. We shall adopt the practice suggested by the Symbols Committee of the Royal Society (1971 and 1975) and use reflectance for the ratio of the reflected to the incident flux.
3. Here, and elsewhere in the literature, for simplicity the irradiance is taken to be the square of the amplitude of the electric vector. In reality, the irradiance is given by the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$. For example, the time average of the Poynting vector for a plane wave is the quantity to which our eyes or a detector would be sensitive and is given by $\langle S \rangle = \frac{1}{2} \epsilon_0 c |\mathcal{E}_0|^2$, where \mathcal{E}_0 is the maximum amplitude of the electric field oscillations. Thus $\langle S \rangle = 1.33 \times 10^{-3} |\mathcal{E}_0|^2 \text{ W m}^{-2}$ in free space (see e.g. ref. 1.1e).

4. The recommended symbol for radiant intensity is I , but we intend to use that symbol for irradiance; the radiant intensity is not frequently used in the text.

PROBLEMS

- 1.1 Express eq. (1.6) in alternative forms such as $\xi = \xi_0 \cos[2\pi(t/T - x/\lambda)] = \xi_0 \cos[k(vt - x)]$. Write down the equation of the wave which is (a) 90° out of phase with these equations and (b) travelling in the negative x direction.
- 1.2 By considering the superposition of two waves which differ in angular frequency and wavenumber by the small amounts $\delta\omega$ and δk , derive eq. (1.8).
- 1.3 Given that the average solar constant is of the order of 700 W m^{-2} in the United Kingdom calculate the corresponding electric field amplitude.
- 1.4 Derive Malus's law, eq. (1.15).
- 1.5 Derive the results of eq. (1.17) for the superposition of two waves by (a) treating the two waves as vectors of magnitudes proportional to the wave amplitude and direction given by the phase of the wave and (b) expressing the waves as complex exponential functions. Show that the energy of the resultant of adding five waves of equal amplitudes ξ_0 and phase constants $0, \pi/4, \pi/2, 3\pi/4, \pi$ is $\xi_0^2(3 + 2\sqrt{2})$. Show also that the resultant energy of superposing an infinite series of waves of amplitudes $\xi_0, \xi_0/2, \xi_0/4, \xi_0/8, \dots$ and phase constants $0, \pi/2, \pi, 3\pi/2, \dots$ is $4\xi_0^2/5$.
- 1.6 Show that the fringe spacing in a Young's slits experiment is given by $\delta y = \lambda x/H$. If the aperture to screen distance is 1.5 m and the wavelength is 632.8 nm, what slit separation is required to give a fringe spacing of 1.2 mm? If a glass plate ($n = 1.5$) of 0.05 mm thickness is placed over one slit, what is the lateral displacement of the fringe system?
- 1.7 Verify eq. (1.25) for interference effects in thin films; explain the coloured appearance of thin films of oil on wet surfaces or of soap bubbles.
- In an experiment to observe fringes of equal inclination, the source emits wavelengths of 500 nm and 600 nm. It is observed that a bright fringe is obtained for both wavelengths when the light passes through the film at $\cos^{-1} 0.7$ to the normal, and that the next such coincidence occurs at an angle $\cos^{-1} 0.8$. Deduce the optical thickness of the film and the orders of the fringes for which the two coincidences occur.
- 1.8 Explain the phenomenon of 'blooming'. Calculate the thickness of a 'blooming' layer of refractive index 2.0 deposited onto a glass substrate of refractive index 1.5 which will give (a) maximum and (b) minimum reflected light. Assume normal incidence and that $\lambda = 500 \text{ nm}$.
- 1.9 What is the maximum number of orders that can be observed using a plane grating of $300 \text{ lines mm}^{-1}$ for normally incident light of wavelength 546 nm? If light of all wavelengths from 400 to 700 nm were used, what wavelengths would be superposed on the 546 nm wavelength in the highest of these orders?
- 1.10 A collimated beam of light from an He-Ne laser ($\lambda = 632.8 \text{ nm}$) falls normally onto a circular aperture of 0.5 mm diameter. A lens of 0.5 m focal length placed just behind the aperture focuses the diffracted light onto a screen. Calculate the distance of the

first dark ring from the centre of the diffraction pattern. Using the same arrangement, what is the minimum separation of another source which could just be resolved if the two sources are 10 m from the aperture?

- 1.11 Show that the functions which describe the energy density within an enclosed cavity, $\rho(\nu, T)$, and the power emitted by a blackbody of unit area, $W(\nu, T)$, are related by the equation

$$W(\nu, T) = \frac{c}{4} \rho(\nu, T)$$

(Hint: consider a small hole made in the side of an otherwise enclosed cavity and determine by integration over a hemisphere of radius c within the cavity how much of the radiation which is inside the cavity will pass through the hole in 1 second.)

- 1.12 Calculate the difference in energy between the Bohr orbits for hydrogen ($Z = 1$) for which $n = 4$ and $n = 2$. What is the wavelength associated with the photon emitted by an electron which undergoes a transition between these levels?
- 1.13 Calculate the minimum frequency of light which will cause the photoemission of electrons from a metal of 2.4 eV work function. What is the maximum kinetic energy of the photoelectrons emitted by light of 300 nm wavelength?

REFERENCES

- 1.1 (a) R. W. Ditchburn, *Light* (2nd edn), Blackie, Glasgow, 1962; (3rd edn), Academic Press, New York, 1976.
 (b) R. S. Longhurst, *Geometrical and Physical Optics* (3rd edn), Longman, Harlow, 1973.
 (c) E. Hecht, *Optics* (2nd edn), Addison-Wesley, Reading, MA, 1987.
 (d) G. R. Fowles, *Introduction to Modern Optics* (2nd edn), Holt, Rinehart & Winston, New York, 1975.
 (e) M. V. Klein and T. C. Furtak, *Optics* (2nd edn), New York, John Wiley, 1986.
 (f) R. D. Guenther, *Modern Optics*, John Wiley, Toronto, 1990.
- 1.2 R. W. Ditchburn, *op. cit.*, Chapter 3.
- 1.3 *Ibid.*, pp. 137–40.
- 1.4 G. R. Fowles, *op. cit.*, pp. 108–19.
- 1.5 See ref. 1.1c, Chapter 5.
- 1.6 See ref. 1.1c, Section 6.3.
- 1.7 D. Marcuse and S. E. Miller, 'Analysis of a tubular gas lens', *Bell Syst. Tech. J.*, **43**, 1159, 1965.
- 1.8 M. Young, *Optics and Lasers – An Engineering Physics Approach*, Springer-Verlag, Berlin, 1977, Chapter 2.

Elements of solid state physics

To understand the operation of many of the optoelectronic devices discussed in later chapters, we need at least an appreciation of the solid state physics of homogeneous materials and of the junctions between different materials. Accordingly, we now examine the mechanisms by which current flows in a solid, why some materials are good conductors of electricity and others are poor conductors, and why the conductivity of a semiconductor varies with temperature, with the concentration of impurities it contains, and with exposure to light.

One of the physical models which has been quite successful in explaining these and related phenomena is the *energy band* model of solids. Before describing this, however, we shall review some relevant concepts from quantum physics. As in section 1.4 it is assumed that the reader is familiar with much of the basic physics of isolated atoms including the description of atomic energy states in terms of four quantum numbers and the application of the Pauli exclusion principle leading to the electron configuration of atoms (ref. 2.1).

Because of space limitations, several of the equations presented in later sections will not be derived in detail. The reader will be guided through their derivation, however, via the problems and the references presented at the end of the chapter.

2.1

Review of some quantum mechanical concepts

In Chapter 1 we saw that light displays a dual nature, that of particle and wave. Now the energy of a light particle, the photon, can be written as $E = h\nu$, where ν is the frequency of the wave associated with the photon. As the rest mass of the photon is zero, its momentum p can be written as

$$p = E/c$$

Therefore

$$p = h\nu/c = h/\lambda \quad (2.1)$$

where λ is the wavelength associated with the photon. This equation, known as the *de Broglie relation*, can be applied quite generally to particles such as electrons, neutrons and atoms, which also exhibit a dual nature as demonstrated, for example, by electron and neutron diffraction.

The question as to what is meant by the wave to be associated with a particle immediately arises. In 1926, Born showed that the wave amplitude is related to the probability of

locating the particle in a given region of space. More specifically, in quantum mechanical problems, as we shall see below, we attempt to find a quantity Ψ called the *wavefunction*, which while having no direct physical meaning itself, is defined in such a way that the probability of finding the particle in the region of space between x and $x + dx$, y and $y + dy$, z and $z + dz$ is given by $\Psi^*\Psi \, dx \, dy \, dz$, where Ψ^* is the complex conjugate of Ψ (Ψ may be complex, i.e. involve the square root of -1). Clearly if the particle exists then

$$\int_{-\infty}^{\infty} \Psi^*\Psi \, dx \, dy \, dz = 1 \tag{2.2}$$

$\Psi^*\Psi$ or $|\Psi|^2$ is called the probability density function.

A real particle, which is localized, cannot be described by a wave equation of the form of eq. (1.6) say, which is of infinite extent. Rather, the wave description of a particle is given in terms of a wave packet (see section 1.2), which is the sum of individual waves with varying amplitudes and frequencies. These waves interfere destructively except for a certain region in space where the probability of finding the particle is high (Fig. 2.1). The speed of the particle v is the same as the group velocity of the wave packet, that is from eq. (1.8)

$$v = v_g = \frac{\partial \omega}{\partial k}$$

The probability of finding the particle is greatest at the centre of the wave packet, $x = 0$ in Fig. 2.1; however, there is a smaller but finite probability of finding it anywhere in the region Δx . Now it may be shown from the Fourier integral (ref. 2.2) that the narrower is Δx , that is the more accurately the position of the particle is known, the wider must be the range

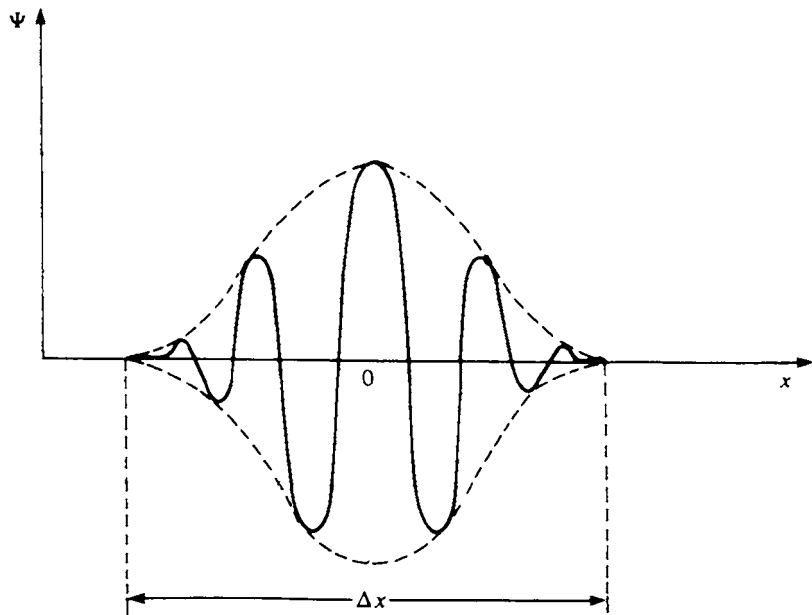


FIG. 2.1 Wave packet of length Δx .

of wavelengths $\Delta\lambda$ comprising the wave packet. From eq. (2.1) this implies a greater uncertainty, Δp , in the momentum of the particle. Heisenberg showed that in any simultaneous measurement of the position and momentum of a particle, the uncertainties in the two measured quantities are related by an equation of the form

$$\Delta x \Delta p \geq \hbar/2 \quad (2.3)$$

Equation (2.3) is called Heisenberg's *uncertainty principle*. This implies that it is impossible to describe with absolute precision events involving individual particles: we can only talk of the *probability* of finding a particle at a certain position at a given instant of time.

2.1.1 Schrödinger equation

In general, the Schrödinger equation includes both space and time dependencies and is of the form

$$\frac{\hbar^2}{2m} \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} \right) - V\Psi = -i\hbar \frac{\partial \Psi}{\partial t} \quad (2.4)$$

where $i = \sqrt{-1}$, V is the potential energy of the particle and Ψ the wavefunction, both of which may depend on position and time. That is, we have $\Psi = \Psi(x, y, z, t)$ and $V = V(x, y, z, t)$. Equation (2.4) is often referred to as the time-dependent Schrödinger equation. Unfortunately, eq. (2.4) is quite complicated and can only be solved in a limited number of cases. For our purposes, it is sufficient to consider situations in which the potential energy of the particle does not depend explicitly upon time. The forces that act upon it, and hence V , then vary only with the position of the particle. In this case, eq. (2.4) reduces to the time-independent form, which, in one dimension, may be written as

$$\frac{d^2 \psi(x)}{dx^2} + \frac{2m}{\hbar^2} [E - V(x)] \psi(x) = 0 \quad (2.4a)$$

where ψ is the time-independent wavefunction and E is the total energy of the particle.

2.1.1.1 Potential well

It is quite difficult to find solutions of the Schrödinger equation for most realistic potential distributions. However, there are several physical situations, for example a free electron trapped in a metal or charge carriers trapped by the potential barriers of a double heterojunction (section 5.10.2.3), which can be approximated by an electron in an infinitely deep, one-dimensional potential well. The situation is illustrated in Fig. 2.2, where for simplicity we have taken $V(x) = 0$ except at the boundaries where $V(x)$ is infinitely large. That is, the boundary conditions are

$$V(x) = 0 \quad 0 < x < L \quad (2.5)$$

and

$$V(x) = \infty \quad x \leq 0, x \geq L \quad (2.5a)$$

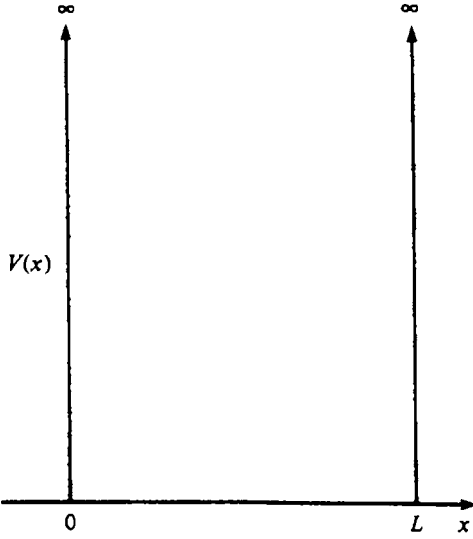


FIG. 2.2 One-dimensional potential well of infinite depth.

Inside the potential well, eq. (2.4a) becomes

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2} E \psi(x) = 0 \quad (2.6)$$

which is the wave equation for a free particle in a region where $V(x) = 0$.

A possible solution to eq. (2.6) is

$$\psi(x) = A \sin kx + B \cos kx \quad (2.7)$$

where A and B are constants and

$$k^2 = 2mE/\hbar^2 \quad (2.8)$$

Applying the boundary conditions given by eqs (2.5) we see that $\psi(x)$ must be zero at the boundaries of the well. Otherwise, there would be a non-zero value of $|\psi|^2$ outside the well, which is impossible because a particle cannot penetrate an infinitely high potential barrier. Thus as $\psi(x) = 0$ at $x = 0$, B must be zero, and as $\psi(x) = 0$ at $x = L$, k must be defined so that $\sin kx$ is zero at $x = L$. That is, kL must be an integral multiple of π . We can therefore write

$$\psi(x) = A \sin kx \quad (2.9)$$

where

$$k = n\pi/L \quad n = 1, 2, 3, \dots \quad (2.10)$$

Substituting for k from eq. (2.8) we have

$$\left(\frac{2mE_n}{\hbar^2} \right)^{1/2} = \frac{n\pi}{L}$$

or

$$E_n = \frac{n^2 h^2}{8mL^2} \quad (2.11)$$

Thus for each value of n the particle energy is described by eq. (2.11). We note that the total energy is *quantized*; n is a *quantum number*.

The value of A in eq. (2.9) can be obtained by using the *normalization condition* which is expressed by eq. (2.2). We find that $A = \sqrt{2/L}$ and hence

$$\psi_n = \left(\frac{2}{L}\right)^{1/2} \sin \frac{n\pi x}{L} \quad (2.12)$$

The wavefunctions ψ_n and the corresponding energies E_n , which are often called *eigenfunctions* and *energy eigenvalues* respectively, describe the quantum state of the particle. The forms of ψ_n and the probability density $|\psi_n|^2$ are shown in Figs 2.3(a) and (b) respectively for the first three quantum states ($n = 1, 2, 3$).

This solution for a one-dimensional potential well can be extended quite easily to the more realistic case of three dimensions, where, assuming that the sides of the potential well are all the same length, the eigenfunctions are given by

$$\psi_n = (8/L)^{1/2} \sin k_1 x \sin k_2 y \sin k_3 z$$

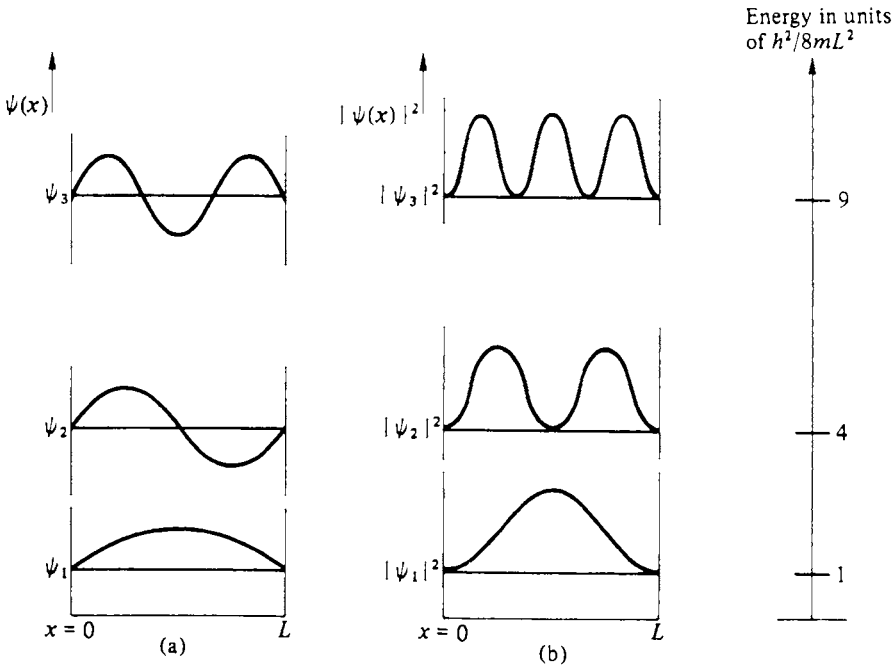


FIG. 2.3 Ground state and first two excited states of an electron in a potential well: (a) the electron wavefunctions and (b) the corresponding probability density functions. The energies of these three states are shown on the right.

where

$$k_1 = n_1\pi/L \quad k_2 = n_2\pi/L \quad k_3 = n_3\pi/L$$

The energy eigenvalues are still given by eq. (2.11), that is

$$E_n = \frac{n^2 h^2}{8mL^2}$$

but now we have

$$n^2 = n_1^2 + n_2^2 + n_3^2 \quad (2.13)$$

We see that each energy state is defined by a set of three quantum numbers, each of which yields a different eigenfunction ψ . From other evidence, it is found that each electron eigenfunction has one of two states of opposite spin sign associated with it. These two states are described in terms of the quantum number m_s , which can take either of the two values $+\frac{1}{2}$ or $-\frac{1}{2}$. Thus four quantum numbers (i.e. n_1, n_2, n_3, m_s) are needed to define a quantum state completely. In fact the need for four quantum numbers is general in any three-dimensional potential, although each potential will give rise to its own unique set of quantum numbers, and these may well describe the quantization of quite different physical parameters. For example, in a hydrogen atom an electron may be considered to be moving under the influence of a single positive point charge. Its quantum state is characterized by the set of four numbers n, l, m_l and m_s . Here, n is the *principal* quantum number, mentioned in section 1.4, which determines the electron energy, l is the *orbital angular momentum* quantum number, which determines the angular momentum of the electron, while m_l is the *magnetic* quantum number, which determines the orientation of the angular momentum vector. Only the spin quantum number is the same in these two cases.

Whilst solutions of the Schrödinger equation tell us what states are available to a single electron, we still need to know which of these states are occupied, especially if we have a situation where a number of electrons are involved. At first sight we might expect that all the electrons would occupy the lowest energy states possible. The implications of this are that in a multielectron atom, for example, the electrons would all tend to occupy the same energy level. Since the chemical properties of an atom are determined by the types of levels the electrons occupy, this assumption would give rise to the atoms of the various elements all having very similar chemical properties. Manifestly this is not the case! In fact electrons are found to obey the Pauli exclusion principle. This forbids any two electrons having the same set of quantum numbers. Thus if we have several electrons (assumed to be non-interacting for simplicity) within a three-dimensional potential well, not more than two of them can have the same values for n_1, n_2 and n_3 , and in that case they must then have different values of m_s .

2.2

Energy bands in solids

As isolated atoms are brought together to form a solid various interactions occur between neighbouring atoms. The forces of attraction and repulsion between atoms find a balance

at the proper interatomic spacing for the crystal. In the process important changes occur in the electron energy levels, which result in the varied electrical properties of different classes of solids.

Qualitatively, we can see that as the atoms are brought closer together the application of the Pauli principle becomes important. When atoms are isolated, as in a gas, there is no interaction of the electron wavefunctions: each atom can have its electrons in identical energy levels. As the interatomic spacing decreases, however, the electron wavefunctions begin to overlap and, to avoid violating the Pauli principle, there is a splitting of the discrete energy levels of the isolated atoms into new levels belonging to the collection of atoms as a whole. In a solid many atoms are brought together so that the split energy levels form a set of *bands* of very closely spaced levels with forbidden energy gaps between them, as illustrated in Fig. 2.4. The lower energy bands are occupied by electrons first; those energy bands which are completely occupied (i.e. full) are not important, in general, in determining the electrical properties of the solid. On the other hand, the electrons in the higher energy bands of the solid are extremely important in determining many of the physical properties of the solid. In particular the two highest energy bands, called the *valence* and *conduction* bands, are of crucial importance in this respect, as is the forbidden energy region between them which is referred to as the *energy gap*, E_g . In different solids the valence band might be completely filled, nearly filled or only half filled with electrons, while the conduction band is never more than slightly filled. The extent to which these bands are, or are not, occupied and the size of the energy gap determines the nature of a given solid.

We may further reinforce our model by considering that in an ideal crystalline solid the atoms are arranged in a perfectly periodic array. The potential experienced by an electron in the solid is correspondingly spatially periodic, so that, after a distance in the crystal equal

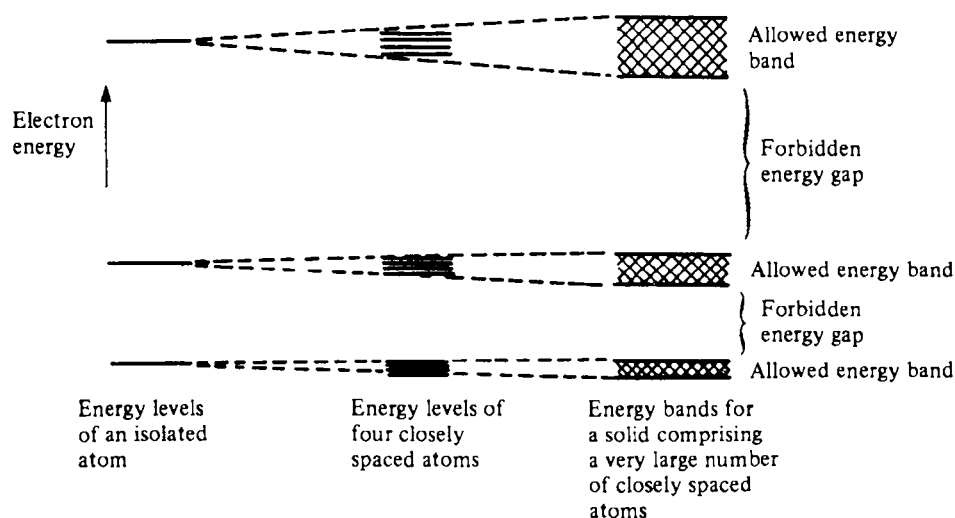


FIG. 2.4 Schematic representation of how the energy levels of interacting atoms form energy bands in solids.

to the lattice spacing, the potential V repeats itself, that is

$$V(x) = V(x + a) = V(x + 2a) = \dots$$

where a is the periodicity of the lattice.

We could now attempt to apply the Schrödinger equation to the problem of electron motion within this system. This is difficult unless we choose a very simple array of atoms. One such approach, the Kronig–Penney model (ref. 2.3), considers a single electron moving within a one-dimensional line of atoms represented by a periodic array of rectangular potential wells. Application of the Schrödinger equation in this case shows that not every value of electron energy is allowed. In fact we find that there are ranges of allowed energies separated by ranges of disallowed energies; that is, the electrons in a solid can occupy certain bands of energy levels which are separated by forbidden energy gaps.

The discontinuities between allowed and forbidden energy values occur at values of the wavevector k given by $k = \pm n\pi/a$, where n is an integer (see Problem 2.6). The $E-k$ curve then takes the appearance shown in Fig. 2.5(a), in contrast to the smooth dashed curve which is the $E-k$ curve for a free electron ($V=0$) – see eq. (2.8). The discontinuities arise from the interaction of the electrons with the periodic potential V ; the corresponding energy band arrangement is shown to the right of Fig. 2.5(a).

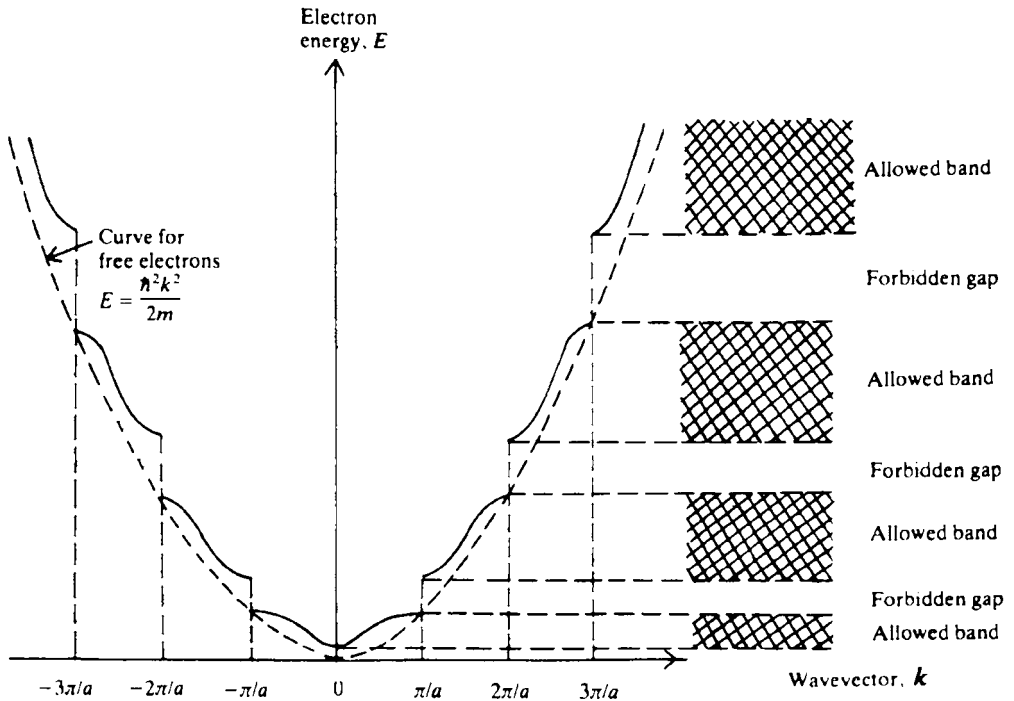


FIG. 2.5(a) Relationship of $E-k$ for electrons subjected to the potential distribution of the Kronig–Penney model and the corresponding energy band structure. The $E-k$ relationship for free electrons is shown for comparison.

The region in Fig. 2.5(a) for which $-\pi/a < k \leq \pi/a$ is called the *first Brillouin zone*. It is often convenient to redraw Fig. 2.5(a) by translating the segments of the $E-k$ curve so that they all lie within this range. This is shown in Fig. 2.5(b). (There is more to this procedure than simply pictorial convenience, as it has a theoretical justification as well, see ref. 2.1f, section 2.2.) In three dimensions the situation is obviously more complicated, but it turns out that diagrams similar to Fig. 2.5(b) can be drawn corresponding to different directions in the crystal. The first Brillouin zone in general has a complicated shape which depends upon the crystal structure being considered. However, its boundaries still lie close to π/a , although the parameter a has now to be interpreted in terms of the crystal unit cell dimensions.

2.2.1 Conductors, semiconductors and insulators

In real crystals the $E-k$ relationship is much more complicated as we can see from Fig. 2.6, which shows the relationships for silicon and gallium arsenide. These depend on the orientation of the electron wavevector with respect to the crystallographic axes, since interatomic distances and the internal potential energy distribution also depend on direction in the crystal.

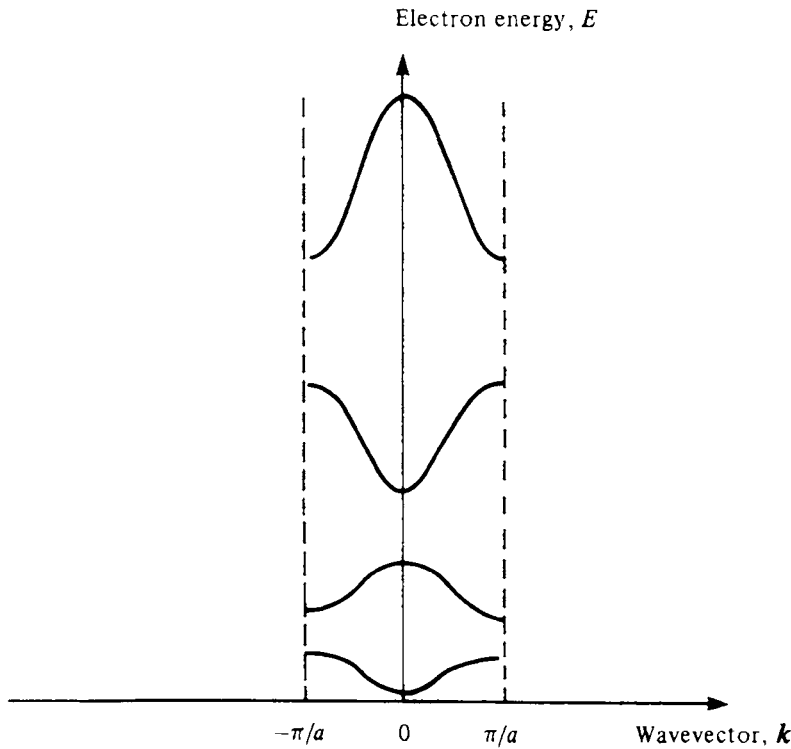


FIG. 2.5(b) Reduced zone representation of the $E-k$ relationship shown in Fig. 2.5(a). This representation is constructed by translating the segments of the $E-k$ curve so that they all lie between $k = -\pi/a$ and $k = +\pi/a$, which comprises the first Brillouin zone.

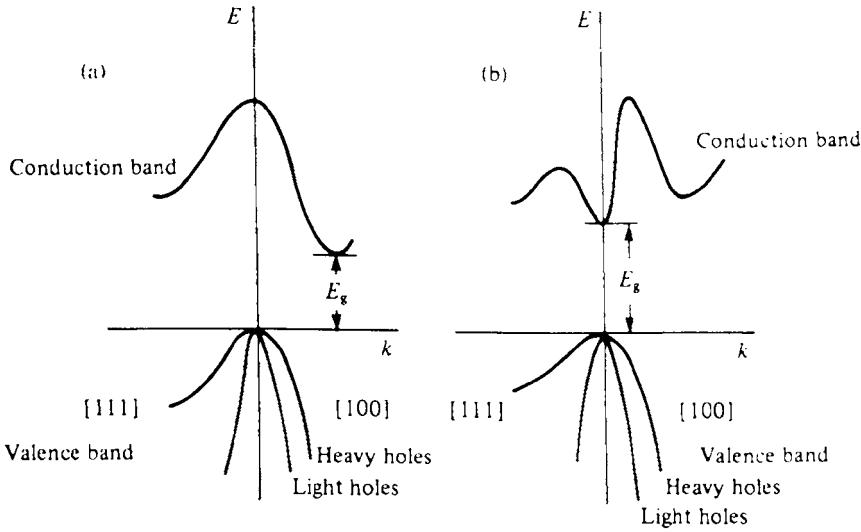


FIG. 2.6 Relationship of E - k for real solids: (a) silicon (which has an indirect bandgap) and (b) gallium arsenide (which has a direct bandgap). The figure shows the conduction and valence bands and the energy gap E_g between them. Note that (i) k is specified in different crystallographic directions to the left and right, and (ii) there are holes present with different effective masses (sections 2.2.1 and 2.3).

However, the basic effect is still energy band formation. One point that arises is that the maximum of the valence band does not always occur at the same k value as the minimum of the conduction band. We speak of a *direct* bandgap semiconductor when they do and an *indirect* bandgap semiconductor when they do not. Thus silicon has an indirect bandgap whilst gallium arsenide has a direct bandgap, which, as we shall see, leads to differing optical properties. For some purposes, however, we may regard the conduction and valence bands as being similar to the bands shown in Fig. 2.5(a).

We may now give a qualitative explanation of why the electrical conductivities of various types of solid are different in terms of the energy band model. The differences arise from the extent to which the energy levels in the valence and conduction bands are filled by electrons. The electrons occupy the allowed states (energy levels) in the energy bands, starting from the lowest, until they are all accommodated (one electron per state). If the electrons in the solid respond to an externally applied magnetic field, and thereby contribute to the conductivity, they will acquire energy from the field and move to higher energy levels; that is, the field accelerates the electrons and therefore increases their energies. This can only occur if there are unoccupied, higher energy levels immediately available to the electrons, either within the same energy band, or in the next higher band providing that the energy gap is very small.

In *metallic conductors*, the uppermost occupied band is only partially filled as shown in Fig. 2.7(a), or there is band overlap (Fig. 2.7b), and electrons can gain energy from an external field quite easily, resulting in high conductivity.

In *insulators* the upper occupied band, the valence band, is completely filled with electrons

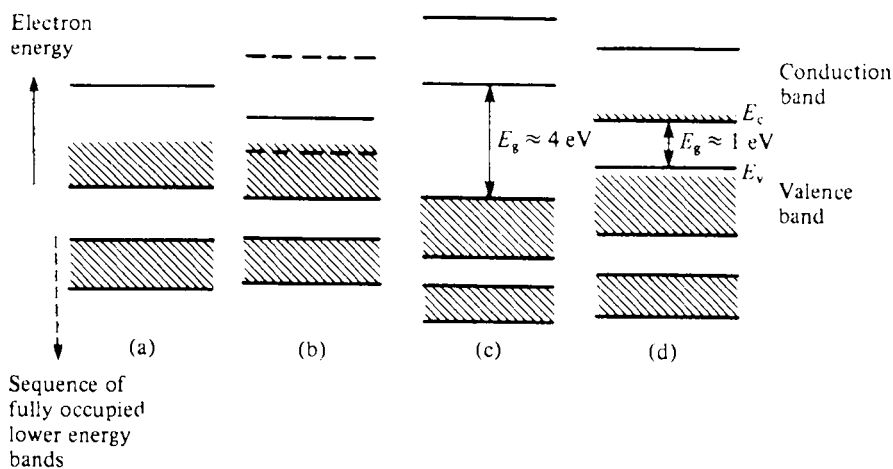


FIG. 2.7 Schematic representation of the energy bands in various materials: (a) a metal with partially filled valence band, e.g. monovalent metals; (b) a metal with two overlapping partially filled bands, e.g. divalent metals; (c) an insulator; and (d) an intrinsic semiconductor. In this, and succeeding diagrams, shading is used to signify occupied electron energy levels.

as shown in Fig. 2.7(c). The nearest empty states are in the conduction band, but these are separated from the valence band by a large energy gap, large, that is, compared with the average thermal energy of the electrons, kT . At room temperature kT is about $1/40$ eV, while typically $E_g \approx 4$ eV. There are, therefore, very few electrons that can respond to a field, and virtually no conduction occurs.

A similar situation arises in intrinsic semiconductors where at low temperatures the valence band is full and the conduction band is empty. In this case, however, the energy gap is sufficiently small (about 1 eV) that some electrons are excited across it at higher temperatures. Clearly, the electrons excited into the conduction band can contribute to current flow. Similarly, as there are now vacant states in the valence band (Fig. 2.7d), the electrons in that band can also respond to an applied field and contribute to the current flow. It is difficult to evaluate this contribution in terms of electron movement, however, as there are a very large number of electrons and a small number of unoccupied states in the band. It turns out that the contribution to the current flow from all the electrons in the nearly full valence band is the same as that which would arise from the presence of a small number of fictitious *positive* charge carriers called *holes* in an otherwise empty band. The number of holes, in fact, is simply equal to the number of empty states in the valence band. Indeed, for many purposes we may regard a hole as an unfilled state. Holes can be regarded as particles which behave in general in a similar way to electrons, apart from having a charge of opposite sign.

The concept of how holes arise in a solid is most easily illustrated by considering the energy band scheme of an intrinsic semiconductor. At any temperature above absolute zero some electrons will be excited from that valence band to the conduction band as a result of their thermal energy. When electrons make such transitions, as illustrated in Fig. 2.8(a), empty states are left in the valence band and we say that *electron-hole* pairs have been created.

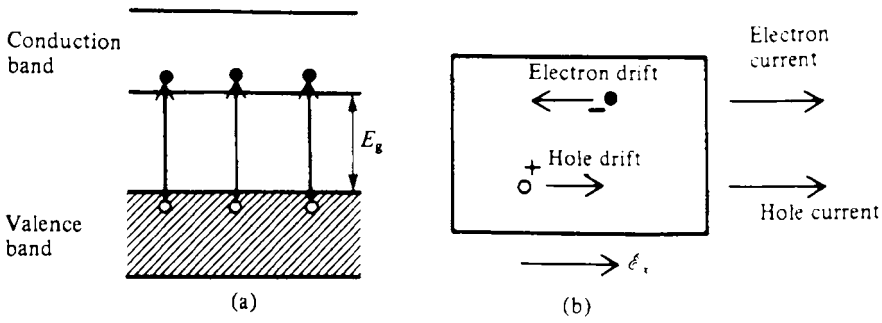


FIG. 2.8 Creation of electron–hole pairs by the thermal excitation of electrons from the valence band to the conduction band (a) and electron and hole drift in a semiconductor on application of a field E_x (b).

If an electric field is now applied to such a solid, electrons in the conduction band and holes in the valence band drift in opposite directions as shown in Fig. 2.8(b) and both thus contribute to the current because of their opposite charges.

2.3 Electrical conductivity

The charge carriers in a solid are in constant thermal motion. In the absence of external fields this motion is quite random resulting from the carriers being scattered by collisions with lattice atoms, impurities and crystal defects. As the motion is random there is no net displacement of charge in any direction and hence no net current flow.

If an electric field E_x is applied then the individual electrons will experience a force $-eE_x$ (where we have taken e to represent the *magnitude* of the electronic charge). While this force may be insufficient to alter the random path of an individual electron appreciably, the effect when averaged over all the electrons is to generate a net motion of the electrons in the negative x direction. If the total momentum of the group of electrons in the x direction is p_x then we may write

$$-neE_x = \frac{dp_x}{dt} \tag{2.14}$$

where n is the electron density. Equation (2.14) indicates a steady acceleration of the electrons in the x direction. However, they lose momentum in collisions with defects in the crystal lattice, such as lattice vibrations (or phonons) and impurities, so that a steady state is reached. In the steady state the rate at which momentum is acquired by the electrons due to the acceleration caused by the electric field is balanced by the rate of loss of momentum due to collisions. Consequently the group of electrons acquires a constant drift velocity v_{1x} . Now if the electron collisions are random the probability of each electron colliding at any instant of time is constant. Thus if we have a group of n electrons at the instant of time $t = 0$, and at time t

$n(t)$ of these have not undergone a collision, then we may write

$$-\frac{dn(t)}{dt} = \frac{1}{\tau} n(t) \quad (2.15)$$

where τ is a constant of proportionality.

The solution to eq. (2.15) is $n(t) = n_0 \exp(-t/\tau)$. We see that τ , which is called the *mean free time* or *relaxation time*, is the mean time between scattering events, and that the probability that an electron has a collision in the time interval dt is dt/τ . Thus the change in momentum due to collisions in time dt is $dp_x = -p_x dt/\tau$.

The rate of loss of momentum due to collisions, which may be regarded as generating a decelerating or resistive force, is then $dp_x/dt = -p_x/\tau$. Hence in the steady state, from eq. (2.14) we have

$$-\frac{p_x}{\tau} - ne\mathcal{E}_x = 0 \quad (2.16)$$

The average momentum per electron, \bar{p}_x , is p_x/n , which from eq. (2.16) is given by

$$\bar{p}_x = -e\tau\mathcal{E}_x \quad (2.17)$$

As expected, eq. (2.17) indicates that the electrons on average have a constant net drift velocity of

$$v_D = \frac{\bar{p}_x}{m_e^*} = \frac{-e\tau\mathcal{E}_x}{m_e^*} \quad (2.18)$$

where m_e^* is the effective mass of the electron (see below).

As the electron concentration is n , the current density resulting from this drift is

$$J = -nev_D \quad (2.19)$$

Hence combining eqs (2.18) and (2.19) we have

$$J = \frac{ne^2\tau}{m_e^*} \mathcal{E}_x \quad (2.20)$$

or

$$J = \sigma \mathcal{E}_x \quad (2.21)$$

Equation (2.21) is, of course, Ohm's law and σ , the electrical conductivity, is given by

$$\sigma = \frac{ne^2\tau}{m_e^*} = ne\mu_e \quad (2.22)$$

The quantity $\mu_e (= e\tau/m_e^*)$ in eq. (2.22) is called the electron *mobility*. It describes the ease with which the electrons drift in the material and it is a very important quantity in characterizing semiconductor materials.

The two basic types of scattering that influence electron and hole mobilities are known as lattice scattering and impurity scattering. In the former a carrier moving through the crystal

encounters atoms which are out of their normal positions because of thermal vibrations. It is to be expected that the frequency of such events would rise as the temperature rises, leading to a decrease in the mobility with increasing temperature. On the other hand, scattering from lattice defects such as ionized impurities (including dopant atoms, see section 2.4.2) should be relatively independent of temperature. Thus we would expect lattice scattering to dominate at high temperatures and impurity scattering to dominate at low temperatures.

The mobility, as we can see from eq. (2.18), can be defined as the drift velocity per unit electric field. Equation (2.20) can be rewritten, using the mobility, as

$$J = ne\mu_e \mathcal{E}_x \quad (2.23)$$

In semiconductors we must include current flow due to both electrons and holes and the total current density is

$$J = (ne\mu_e + pe\mu_h)\mathcal{E}_x = \sigma \mathcal{E}_x \quad (2.24)$$

where μ_h is the mobility of the holes, and is given by $\mu_h = e\tau/m_h^*$, where m_h^* is the *effective mass* of the holes.

The concept of effective mass arises when we consider the motion of an electron through a crystal when a field is applied. The situation is not governed simply by the laws of classical physics, because the electron is influenced not only by the external field but also by an internal field, produced by the other electrons, and by the periodic potential of the atoms in the crystal. Despite this we may use Newton's law to evaluate the acceleration of the electron providing we accept that the electron will exhibit an effective mass m_e^* which is different from the mass m of a free electron in vacuum. The value of the effective mass depends on the energy of the electron within an energy band and is given by (ref. 2.4)

$$m_e^* = \hbar^2 \left(\frac{d^2 E}{dk^2} \right)^{-1} \quad (2.25)$$

For an electron at the top of an energy band it can be seen from Fig. 2.5(a) that $d^2 E/dk^2$, and hence also the effective mass, is *negative*. Such an electron will be accelerated by a field in the reverse direction of that expected for a negatively charged electron. The existence of electrons with negative effective masses is closely linked with the concept of holes, which have positive effective masses, m_h^* (ref. 2.5).

As mentioned at the start of section 2.2.1 the $E-k$ relationships depend strongly on direction within the crystal. Hence it might appear from eq. (2.25) that we should have an anisotropic effective mass. In general, however, material parameters such as electrical conductivity that depend on effective mass are found to be isotropic. This is because there are often several minima in the $E-k$ diagram, each giving rise to an anisotropic effective mass, but each being differently orientated so that when an average is taken an isotropic result is obtained. In addition, the use of effective mass in different contexts (e.g. conductivity and density of states (section 2.5)) necessitates taking the average in different ways resulting in different values for the effective mass. Thus we speak of the 'conductivity effective mass' and the 'density of states effective mass'. In silicon, for example, for electrons these are $0.26m$ and $0.55m$ respectively.

To evaluate conductivities from eq. (2.22), in addition to carrier mobilities we also require

values for the carrier concentrations. This is considered in more detail in section 2.5, where we shall find that carrier concentrations are strongly temperature dependent. As the mobilities, as mentioned above, are also temperature dependent, the conductivity varies with temperature in a quite complex way. This variation, however, is the basis of one method of measuring the excitation energies of the impurities (section 2.4.2) and energy gaps (ref. 2.6).

2.4 Semiconductors

2.4.1 Intrinsic semiconductors

A perfect semiconductor crystal containing no impurities or lattice defects is called an intrinsic semiconductor. In such a material, there are no charge carriers at absolute zero but as the temperature rises electron–hole pairs are generated as explained in section 2.2.1. As the carriers are generated in pairs the concentration n of electrons in the conduction band equals the concentration p of holes in the valence band. Thus we have

$$n = p = n_i$$

where n_i is the intrinsic carrier concentration. The value of n_i varies exponentially with temperature (section 2.5), but at room temperature it is usually not very large. For example, in silicon $n_i \approx 1.61 \times 10^{16} \text{ m}^{-3}$ at room temperature, whereas there are about 10^{29} free electrons per cubic metre in a typical metal. Consequently the conductivity of metals is very much greater than that of intrinsic semiconductors as illustrated in Example 2.1.

EXAMPLE 2.1 Electrical conductivity of metals and semiconductors

We may compare the electrical conductivity of copper and intrinsic silicon from the following data: for copper we have a density of $8.93 \times 10^3 \text{ kg m}^{-3}$, an atomic mass number of 63.54, a mean free time between collisions of $2.6 \times 10^{-14} \text{ s}$ and we assume $m_e^* = m$; for intrinsic silicon we have $n = p = n_i = 1.6 \times 10^{16} \text{ m}^{-3}$, an electron mobility of $0.135 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ and a hole mobility of $0.048 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$.

In copper we assume that the free electron concentration is the same as the number of atoms per unit volume, which we may determine from Avogadro's number to be

$$n = \frac{6 \times 10^{26} \times 8.93 \times 10^3}{63.54} = 8.4 \times 10^{28} \text{ m}^{-3}$$

Then from eq. (2.22)

$$\sigma_{\text{Cu}} = 6.4 \times 10^7 \Omega^{-1} \text{ m}^{-1}$$

For intrinsic silicon, using eq. (2.24) we have

$$\sigma_{\text{Si}} = e n_i (\mu_e + \mu_h) = 1.6 \times 10^{-19} \times 1.6 \times 10^{16} \times (0.135 + 0.048)$$

$$\text{or } \sigma_{\text{Si}} = 4.7 \times 10^{-4} \Omega^{-1} \text{ m}^{-1}$$

As at a given temperature there is a steady state carrier concentration, there must be a *recombination* of electron–hole pairs at the same rate as that at which the thermal *generation* occurs. Recombination takes place when an electron in the conduction band makes a transition into a vacant state in the valence band. The energy released in the recombination, which is approximately equal to E_g , may be emitted as a photon or given up as heat to the crystal lattice in the form of quantized lattice vibrations, which are called *phonons*, depending on the nature of the recombination mechanism. When a photon is released, the process is called *radiative recombination*. The absence of photon emission indicates a *non-radiative* process, in which lattice phonons are created.

We may distinguish between two types of recombination process which we term ‘band-to-band’ and ‘defect centre’ recombinations. (Some texts refer to these as ‘direct’ and ‘indirect’ transitions respectively. This terminology has not been adopted here to avoid confusion with direct and indirect bandgaps, ref. 2.7.) In the band-to-band process, which is shown in Fig. 2.9(a), an electron in the conduction band makes a transition directly to the valence band to recombine with a hole. In the defect centre process, the recombination takes place via recombination centres or traps. These are energy levels E_r in the forbidden energy gap which are associated with defect states caused by the presence of impurities or lattice imperfections. Any such defect state can act as a recombination centre if it is capable of trapping a carrier of one type and subsequently capturing a carrier of the opposite type, thereby enabling them to recombine.

The precise mechanism of a defect centre recombination event depends on the nature and energy of the defect state. One such process is illustrated in Fig. 2.9(b). In the first step (i) an electron is trapped by the recombination centre, which subsequently captures a hole (ii). When both of these events have occurred the net result is the annihilation of an electron–hole pair leaving the centre ready to participate in another recombination event (iii). The energy released in the recombination is given up as heat to the lattice.

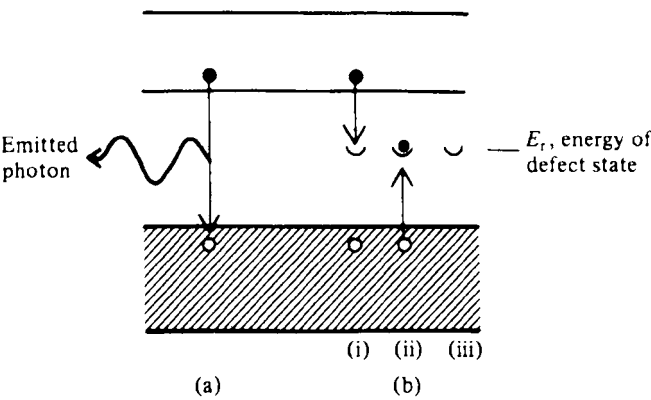


FIG. 2.9 Illustration of (a) band-to-band recombination and (b) recombination via a defect centre. The first step in (b) is (i) the trapping of an electron followed by (ii) hole capture. This results in the annihilation of an electron–hole pair leaving the centre ready to participate in another recombination (iii).

If the thermal generation rate is g_i and the recombination rate is r_i , then, in equilibrium,

$$g_i = r_i \quad (2.26)$$

Both rates are temperature dependent so that if the temperature is raised, g_i increases and a new value of carrier concentration n_i is established such that the higher recombination rate just balances generation.

At any temperature the probability of an electron recombining is proportional to the number of holes present; thus electrons will disappear at a rate proportional to the product of the electron and hole concentrations. Therefore we may write, in general,

$$r_i = Bnp = g_i \quad (2.27)$$

where B is a constant of proportionality which depends on the recombination mechanism taking place (see also section 4.6). For an intrinsic material $n = p = n_i$ and $r_i = Bn_i^2$.

2.4.2 Extrinsic semiconductors

The number of charge carriers in a semiconductor can be vastly increased by introducing appropriate impurities into the crystal lattice. In this process, which is called *doping*, a crystal can be altered so that it has a predominance of either electrons or holes; that is, it can be made either *n*-type (where the *majority* carriers are *negative* electrons and the *minority* carriers are holes) or *p*-type (where the majority carriers are *positive* holes). In doped semiconductors the carrier concentrations are no longer equal and the material is said to be *extrinsic*.

In doping tetravalent elements, for example silicon, impurities from column V of the periodic table such as phosphorus and arsenic or from column III such as boron and indium are used to produce *n*-type and *p*-type semiconductors respectively. The reasons for this are as follows. When intrinsic silicon is doped with phosphorus, for example, the phosphorus atoms are found to occupy atomic sites normally occupied by silicon atoms as shown in Fig. 2.10(a). Since the silicon atoms are tetravalent only four of the five valence atoms of phosphorus are used in forming covalent bonds (ref. 2.8), leaving one electron weakly bound to its parent atom. This electron is easily freed; that is, it can easily be excited into the conduction band. Therefore in the energy band model the energy levels for the 'extra' electrons associated with these impurities lie at E_d , just beneath the conduction band as shown in Fig. 2.10(b). Such impurities are referred to as *donors*, and the energy levels at E_d as *donor* levels, since they donate electrons to the conduction band. The energy required to excite an electron from the donor levels into the conduction band is E_D , which equals $(E_c - E_d)$, where E_c is the energy of the bottom of the conduction band. If, as is frequently the case, we take the energy E_v at the top of the valence band to be zero, then $E_g = E_c$ and then E_D equals $(E_g - E_d)$. At absolute zero the donor levels are all occupied but, because E_D is so small (about 0.04 eV), even at moderately low temperatures most of the electrons are excited into the conduction band, thereby increasing the free electron concentration and the conductivity of the material (see section 2.3).

We can estimate E_D as follows. If a phosphorus impurity atom loses its fifth valence elec-

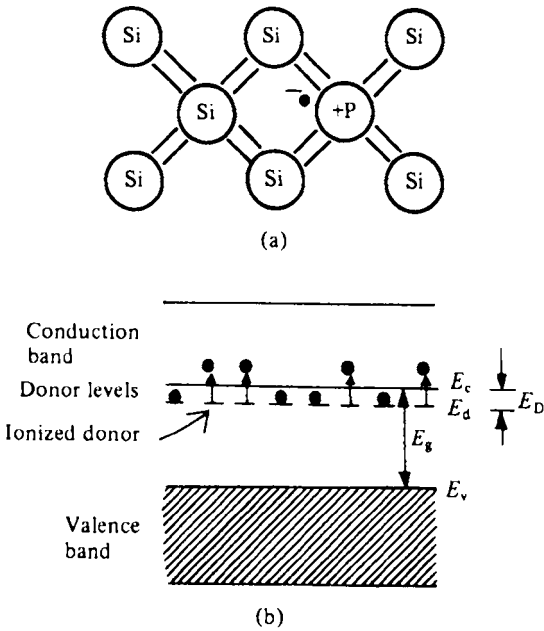


FIG. 2.10 Schematic representation of an n-type semiconductor: (a) the crystal lattice in which a silicon atom has been replaced by a phosphorus impurity atom and (b) the energy levels of the donor impurity atoms (the diagram shows that some of the impurities have ‘donated’ an electron to the conduction band).

tron it is left with a net positive charge of $+e$ (the impurity is said to have been ionized). It can be imagined, therefore, that this electron is bound to its parent atom in a situation which is similar to that found in the hydrogen atom, where a charge of $+e$ binds an electron to the nucleus. The ionization energy of the hydrogen atom is 13.6 eV (see Example 1.5), but in the case under discussion here there are two important differences arising from the fact that the electron moves in a solid. First we must use the effective mass m_e^* rather than the free electron mass. Secondly the relative permittivity of the semiconductor must be included in the derivation of the electron energy levels. This is because the electron orbit is large enough to embrace a significant number of silicon atoms so that the electron may be considered to be moving in a dielectric medium of relative permittivity ϵ_r . Therefore, from eq. (1.43), the excitation energy E_D is given by

$$E_D = 13.6 \frac{m_e^*}{m} \left(\frac{1}{\epsilon_r} \right)^2 \text{ eV} \tag{2.28}$$

Suppose, on the other hand, that silicon is doped with boron. Again it is found that the impurity atoms occupy sites normally occupied by silicon atoms as shown in Fig. 2.11(a). In this case, however, there is one electron too few to complete the covalent bonding. At temperatures above absolute zero an electron from a neighbouring silicon atom can move to the impurity to complete the bonding there but, by so doing, it leaves a vacant state in the valence band thereby creating an additional hole. For this reason the trivalent impuri-

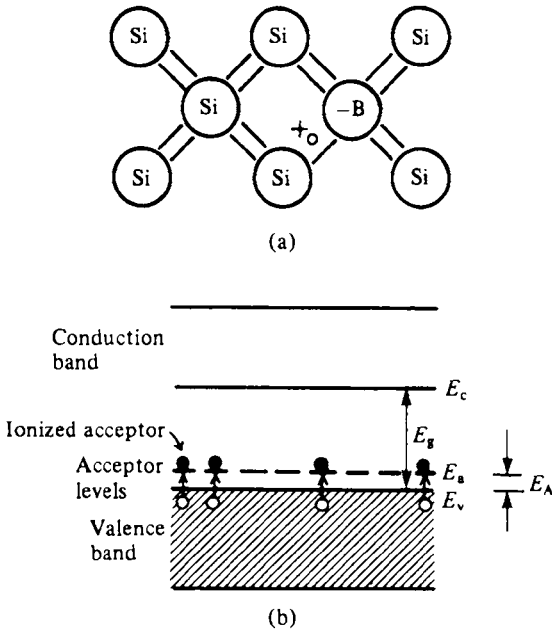


FIG. 2.11 Schematic representation of a p-type semiconductor: (a) the crystal lattice with trivalent impurities (e.g. boron) and (b) the energy levels of the acceptor impurity atoms (some impurities have 'accepted' electrons from the valence band).

ties are referred to as *acceptors* as they accept electrons excited from the valence band. It is convenient to regard this situation as a negatively ionized acceptor atom with a positive hole orbiting around it analogous to the situation described above. The energy E_A , which equals $E_a - E_v$, required to 'free' the hole from its parent impurity can be estimated as above. An average value of the effective mass of holes in silicon is $0.33m$ and, from eq. (2.28), $E_A = 0.032$ eV. In reality, of course, E_A is the energy required to excite an electron from the valence band to the acceptor energy levels, which lie just above the valence band as illustrated in Fig. 2.11(b).

EXAMPLE 2.2 Ionization energy of donor impurities

We may estimate the energy required to excite electrons from the donor levels to the conduction band in silicon given that $m_e^* = 0.26m$ and the relative permittivity is 11.8.

From eq. (2.28) we then have

$$E_D = 13.6 \times 0.26 \left(\frac{1}{11.8} \right)^2 = 0.025 \text{ eV}$$

(We may compare this value with the following experimental values for various donor impurities in silicon: P 0.045 eV; As 0.05 eV; and Sb 0.04 eV.)

In the compound semiconductor gallium arsenide, gallium is from group III of the periodic table, while arsenic is from group V (such semiconductors are often referred to as group III–V compounds; other members of this group are GaP, InP and AlSb). In this case n- and p-type doping can be accomplished in a number of different ways. For example, if a group II element such as zinc replaces the gallium atoms, the material becomes p-type, whereas if a group VI element such as tellurium replaces the arsenic atoms, it becomes n-type. Group IV elements such as germanium and silicon can produce either p- or n-type material depending on whether the group IV impurity replaces the arsenic or gallium ions respectively.

Similar remarks apply to the group II–VI compound semiconductors, for example CdS and ZnSe, and to the ternary and quaternary semiconductors such as GaInAs and GaInAsP.

2.4.3 Excitons

We have just seen that the introduction of suitable impurities into intrinsic semiconductor material can result in the formation of electron energy levels situated just below the bottom of the conduction band. However, electron energy levels similarly situated can also appear in *intrinsic* material. These arise because the Coulombic attraction of an electron for a hole can result in the two being bound together; such a bound electron–hole pair is called an *exciton*. We may picture the exciton as an electron and a hole orbiting about their common centre of gravity with orbital radii which are inversely proportional to their effective masses as shown in Fig. 2.12. The Bohr model is readily adapted to this situation with the electron mass being replaced by the reduced mass m_r^* of the electron and hole; m_r^* is given by

$$\frac{1}{m_r^*} = \frac{1}{m_e^*} + \frac{1}{m_h^*}$$

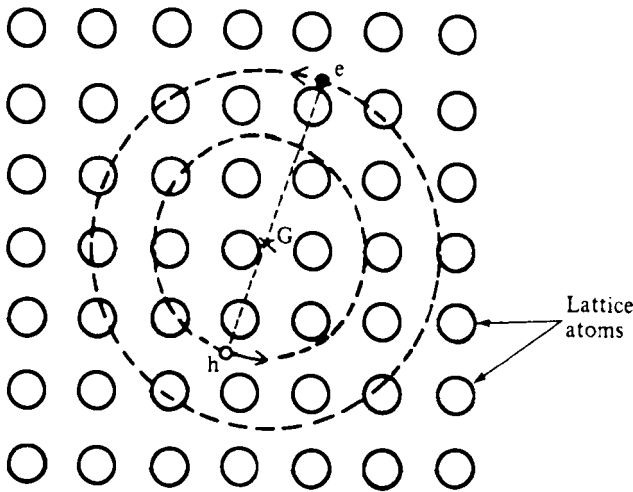


FIG. 2.12 Model of an exciton; the electron e and hole h may be regarded as being bound together and orbiting around their common centre of gravity G with radii which are inversely proportional to their effective masses (we have assumed $m_h^* > m_e^*$).

The binding energy E_c is then obtained by modifying eq. (2.28) to give

$$E_c = 13.6 \frac{m_t^*}{m} \left(\frac{1}{\epsilon_t} \right)^2 \text{ eV} \quad (2.28a)$$

Since m_t^* will be the same order of magnitude as m_c^* and m_h^* , we see that exciton energy levels will be similarly placed to those of donor levels in doped semiconductors.

Excitons may move through a crystalline lattice and thus provide an important means of transferring energy from one point in the material to another. They play an important role in the luminescence of solids and will be discussed further in Chapter 4.

2.5 Carrier concentrations

In calculating semiconductor properties and in analyzing device behaviour it is often necessary to know the carrier concentrations. In metals we can make a fairly good estimate of the free electron concentration by calculating the number of atoms per unit volume (from the density and atomic mass number of the metal) and multiplying by the valency. Similarly in heavily doped semiconductors we can take the majority carrier concentration to be the same as the impurity concentration. This may not be the case at high temperatures when the number of electron-hole pairs generated by electron excitation across the energy gap may be greater than the number of impurities. The situation in near-intrinsic material is not so clear, however, nor is the temperature variation of carrier concentrations immediately obvious.

To calculate the carrier concentration in each energy band we need to know the following parameters:

1. The *distribution of energy states* or levels as a function of energy within the energy band.
2. The *probability* of each of these states being occupied by an electron.

The first of these parameters is given by the *density of states function* $Z(E)$ which may be defined as the number of energy states per unit energy per unit volume. The form of $Z(E)$, which is shown in Fig. 2.13(a), can be derived, for example, from eq. (2.11), which gives the energy levels in a potential well. It is given by (ref. 2.9)

$$Z(E) = \frac{4\pi}{h^3} (2m_e^*)^{3/2} E^{1/2} \quad (2.29)$$

where E is measured relative to the bottom of the band.

The second parameter depends on the fact that electrons obey the Pauli exclusion principle and hence the probability of a particular energy level being occupied at temperature T is given by Fermi-Dirac statistics. This is in contrast to the case of atoms in an ideal gas where the Maxwell-Boltzmann distribution function applies (in fact, in several situations electron behaviour can be approximated by Maxwell-Boltzmann statistics thereby simplifying the mathematics).

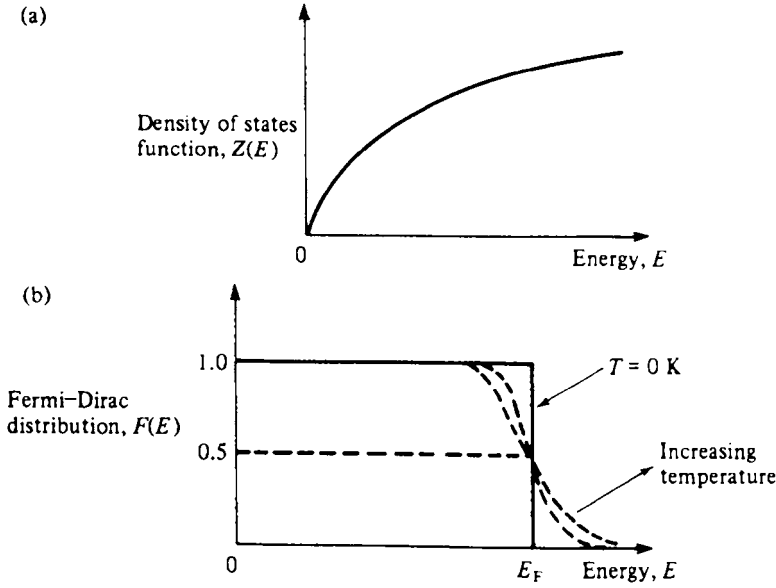


FIG. 2.13 Density of states function (a) and the Fermi-Dirac function at $T = 0 \text{ K}$ and at $T > 0 \text{ K}$ (b).

The Fermi-Dirac distribution function is found to be

$$F(E) = \frac{1}{\exp[(E - E_F)/kT] + 1} \quad (2.30)$$

where E_F is a characteristic energy called the *Fermi energy*.

The distribution function is shown in Fig. 2.13(b). We notice that at 0 K $F(E)$ is unity for energies less than E_F and zero for energies greater than E_F . At any temperature above absolute zero the probability of occupation of the energy level at $E = E_F$ is 0.5, as we can see from eq. (2.30). Figure 2.13(b) also shows that the probability of occupation of states above E_F is finite for $T > 0$ and that there is a corresponding probability that states below E_F are empty. In fact, $F(E + E_F) = 1 - F(E_F - E)$. This makes the Fermi level a natural reference point when calculating electron and hole concentrations. In the case of intrinsic semiconductors the 'tails' in the probability distribution extend into the conduction and valence bands respectively as shown in Fig. 2.14(b); the density of states function and carrier densities are shown in Figs 2.14(a) and (c). As the electron and hole concentrations are equal we expect the Fermi level to lie near to the middle of the energy gap.

In n-type material there are many more electrons in the conduction band than holes in the valence band and we expect the Fermi level to lie near the donor levels. Similarly in p-type material the Fermi level lies near the acceptor levels. The Fermi level, density of states and carrier densities in n-type material are shown in Fig. 2.15. In many cases, for example when considering junctions between dissimilar materials, the energies of the impurity levels can be ignored if we know where the Fermi level lies.

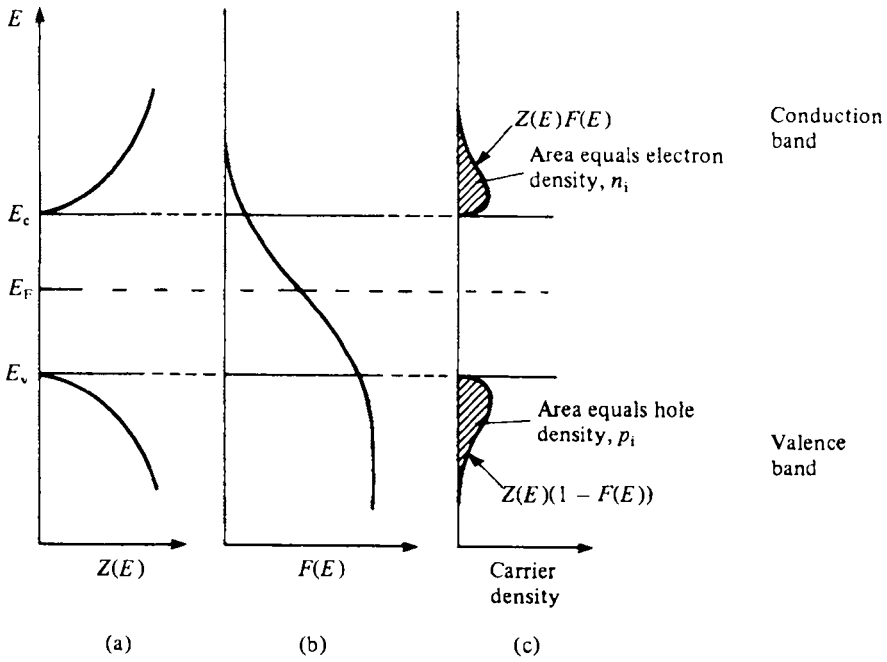


FIG. 2.14 Graphical representation of (a) the density of states, (b) the Fermi-Dirac distribution and (c) carrier densities for an intrinsic semiconductor.

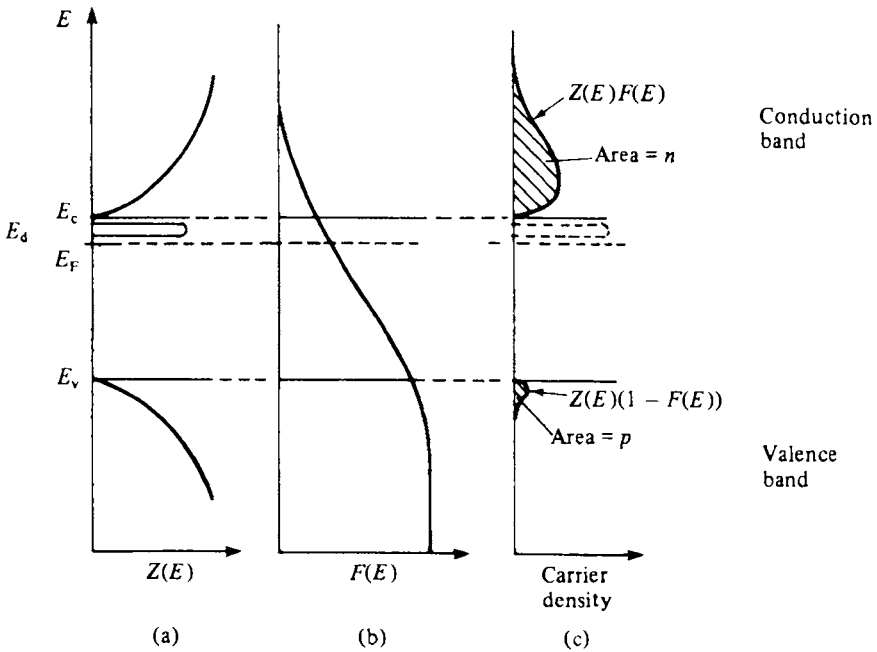


FIG. 2.15 Graphical representation of (a) the density of states, (b) the Fermi-Dirac distribution and (c) carrier densities for an n-type semiconductor.

Returning to the calculation of carrier concentrations we see that this is given by summing the product of the density of states and the occupancy probability over the energy range of interest; that is,

$$n = \int_{\text{energy band}} F(E) Z(E) dE \quad (2.31)$$

Taking first the case of a metal at absolute zero, where $F(E) = 1$ and the upper occupied level is E_F , we have

$$n = \int_0^{E_F} \frac{4\pi}{h^3} (2m_e^*)^{3/2} E^{1/2} dE$$

and hence

$$n = \frac{8\pi}{3h^3} (2m_e^* E_F)^{3/2} \quad (2.32)$$

In fact we often calculate n as described earlier and use eq. (2.32) to calculate E_F ; typical values are 7.0 eV for copper and 11.2 eV for aluminium which are in good agreement with experimental results.

For the case of the electron concentration in the conduction band of a semiconductor we replace E in eq. (2.29) by $(E - E_c)$ so that eq. (2.31) becomes

$$n = \frac{4\pi}{h^3} (2m_e^*)^{3/2} \int_{E_c}^{E_T} \frac{(E - E_c)^{1/2}}{\exp[(E - E_F)/kT] + 1} dE$$

where E_T is the top of the band. Making the substitutions and approximations suggested in Problem 2.14 we have

$$n = N_c \exp \left[- \left(\frac{E_c - E_F}{kT} \right) \right] \quad (2.33)$$

where

$$N_c = 2 \left(\frac{2\pi m_e^* kT}{h^2} \right)^{3/2}$$

N_c is called the *effective density of states* in the conduction band; it is constant at a constant temperature.

EXAMPLE 2.3 Effective density of states in the conduction band

We may calculate the effective density of states for germanium at 300 K given that the appropriate effective mass is $0.55m$.

From eq. (2.33) we have $N_c = 1.03 \times 10^{25} \text{ m}^{-3}$.

By a similar argument, the concentration of holes in the valence band is given by

$$p = \int_{\text{valence band}} [1 - F(E)] Z(E) dE$$

or

$$p = N_v \exp[(E_F - E_v)/kT] \quad (2.34)$$

where

$$N_v = 2 \left(\frac{2\pi m_h^* kT}{h^2} \right)^{3/2}$$

The carrier concentrations predicted by eqs (2.33) and (2.34) are valid for *any* type of semiconductor provided the appropriate Fermi level is used, and Fig. 2.15 shows the carrier distributions, together with the functions $Z(E)$ and $F(E)$, for an n-type semiconductor. We may use these equations to derive an expression for E_F in intrinsic semiconductors where $n = p = n_i$. It is left to the reader to show that

$$E_{Fi} = \frac{1}{2} E_g + \frac{3}{4} kT \ln(m_h^*/m_e^*) \quad (2.35)$$

The second term on the right-hand side of eq. (2.35) is usually very small and, as we supposed above, $E_{Fi} \approx \frac{1}{2} E_g$.

We also note from eqs (2.33) and (2.34) that the product of n and p is constant for a given piece of semiconductor at constant temperature. That is,

$$np = n_i p_i = n_i^2 = N_c N_v \exp(-E_g/kT) \quad (2.36)$$

Equation (2.36) is a very important relationship, for once we know n or p then the other can be determined from n_i^2 ; it is often referred to as the 'law of mass action' for semiconductors.

Equation (2.36) shows that the temperature variation of n_i is essentially exponential, that is $n_i \propto \exp(-E_g/2kT)$. Similarly, for an n-type semiconductor the excitation of electrons from the donor levels to the conduction band is governed by the exponential function $\exp(-E_D/kT)$. As E_D is rather small, we find that at temperatures above about 100 K nearly all the donors are ionized, while at such temperatures the number of electron-hole pairs formed by intrinsic excitation is negligible. This situation continues until the temperature rises somewhat above room temperature in the case of silicon. Thus the variation of electron concentration with temperature has the form shown in Fig. 2.16, the exact details of this variation depending on the doping level.

Finally if we denote the number of electrons having energies between E and $E + dE$ in the conduction band of a semiconductor as $n(E)$ we may write

$$n(E) = \frac{4\pi}{h^3} (2m_e^*)^{3/2} (E - E_c)^{1/2} \exp \left[- \left(\frac{E - E_F}{kT} \right) \right]$$

This function is shown in Fig. 2.14(c) for intrinsic semiconductors, from which it can be seen that $n(E)$ rises fairly rapidly to a maximum and then slowly decreases. It can be shown that the maximum value of $n(E)$ occurs at an energy of $E_c + kT/2$ and that $n(E)$ falls to one-

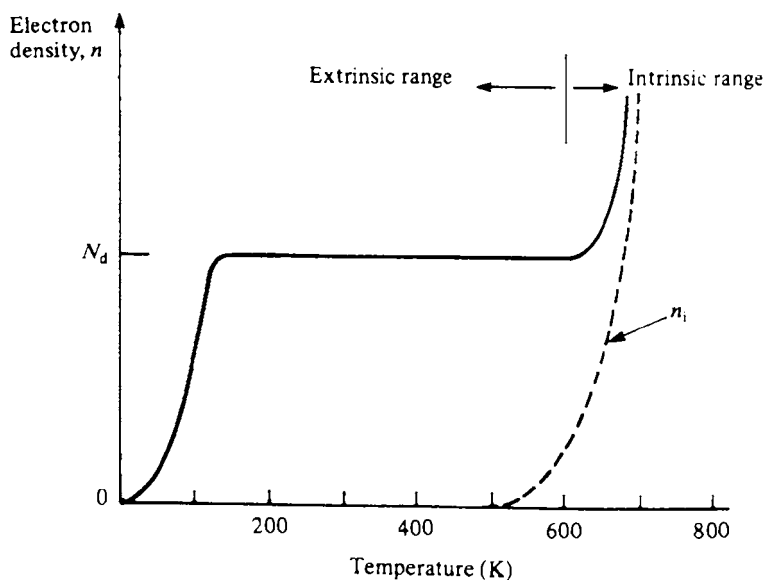


FIG. 2.16 Electron density as a function of temperature in n-type silicon; N_d is the density of donor impurity atoms.

half of this value over an energy range of $1.8kT$ (see Problem 2.20). Thus we see that the electrons within the conduction band are predominantly distributed over an energy range of about $2kT$. Similar comments apply to the distribution of holes in the valence band. We shall return to this point in section 2.9.

2.6 Work function

We saw in section 1.4 that when light of an appropriate frequency falls onto metals (and indeed onto other solids) electrons may be emitted. Similarly, when solids are heated thermionic emission of electrons may occur. The minimum energy required to enable an electron to escape from the surface of the solid, in either case, is called the *work function* ϕ .

A simple model of this situation in the case of a metal is shown in Fig. 2.17(a), which represents an electron trapped in a well of *finite* depth (i.e. one from which it may escape). We see that ϕ is the energy difference between the Fermi level (which corresponds to the highest occupied energy level) and the vacuum level. We consider that when the electron reaches an energy equivalent to the vacuum level it has escaped from the metal. We might compare this situation with the ionization of an atom.

In semiconductors the work function is also defined as the energy difference between the Fermi and vacuum levels (Fig. 2.17b), but in this case it is more usual to use the *electron affinity* χ , defined as the energy difference between the bottom of the conduction band and

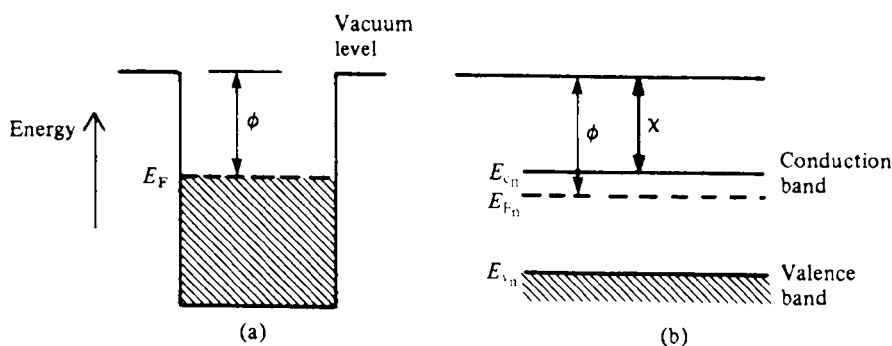


FIG. 2.17 Diagram showing the work function of (a) a metal and (b) an n-type semiconductor.

the vacuum level. In some respects χ has more physical significance than ϕ as, in general, there are no electrons at the Fermi level in semiconductors.

2.7 Excess carriers in semiconductors

Most semiconductor devices operate by the creation of charge carrier concentrations in excess of the thermal equilibrium values given by eqs (2.33) and (2.34). These excess concentrations may be created by *optical excitation* or *carrier injection* via a suitable contact. In optical excitation electron-hole pairs are generated when the energy of the incident photons is sufficient to excite an electron from the valence to the conduction band, that is when $h\nu \geq E_g$. Since additional carriers are thereby created the electrical conductivity of the material will increase. This is the basis of the photoconductive devices described in section 7.3.5.

If a carrier excess is created in this way and the exciting radiation is then suddenly cut off, the carrier concentrations will return gradually to their thermal equilibrium values owing to recombination of the excess carriers. During this time, the recombination rate, being proportional to the increased carrier concentrations, will be higher than the thermal generation rate, assuming that the temperature remains constant.

If we suppose that $\Delta n(t)$ and $\Delta p(t)$ are the excess carrier concentrations at any given instant ($\Delta n(t) = \Delta p(t)$), the net rate at which the excess electron-hole pairs disappear is, from eq. (2.27),

$$-\frac{dn(t)}{dt} = -\frac{d\Delta n(t)}{dt} = B[n + \Delta n(t)][p + \Delta p(t)] - Bn_i^2 \quad (2.37)$$

where the total concentration of electrons at time t is $n(t) = n + \Delta n(t)$, n being the equilibrium concentration (similarly for holes). As $np = n_i^2$, and $\Delta n(t) = \Delta p(t)$, we can rewrite eq. (2.37) as

$$-\frac{d\Delta n(t)}{dt} = B[(n + p)\Delta n(t) + (\Delta n(t))^2]$$

If the excess concentrations are not too high (i.e. we have low level injection) we can ignore the term $(\Delta n(t))^2$. Furthermore, if the material is extrinsic we can usually neglect the minority carrier concentration. Thus, if the material is p-type we have $p \gg n$, and hence we may write

$$-\frac{d\Delta n(t)}{dt} = Bp\Delta n(t)$$

The solution to this equation is

$$\Delta n(t) = \Delta n(0) \exp(-Bpt) = \Delta n(0) \exp(-t/\tau_e) \quad (2.38)$$

where $\Delta n(0)$ is the excess carrier concentration at $t=0$ when the exciting source is switched off and the decay constant, $\tau_e = 1/Bp$, is called the *minority carrier recombination lifetime* (minority, because the calculation is made in terms of the minority carriers). Similarly, the minority carrier lifetime of holes in n-type material is $\tau_h = (Bn)^{-1}$. Physically τ represents the average time an excess carrier remains free before recombining directly. The carrier lifetime for indirect recombination is more complicated than in the above case since it is necessary to include the time required to capture each type of carrier.

EXAMPLE 2.4 Minority carrier lifetimes

We may calculate the minority carrier recombination lifetime for direct recombination in gallium arsenide and silicon for a donor concentration of $5 \times 10^{24} \text{ m}^{-3}$ using the data provided in Appendix 6.

We see above that the hole lifetime in n-type material is given by $\tau_h = 1/Bn$. Hence

$$\text{for GaAs} \quad \tau_h = \frac{1}{7.2 \times 10^{-16} \times 5 \times 10^{24}} = 278 \text{ ps}$$

$$\text{for Si} \quad \tau_h = \frac{1}{1.8 \times 10^{-21} \times 5 \times 10^{24}} = 111 \text{ } \mu\text{s}$$

We thus see that direct, radiative recombination is much more likely in gallium arsenide than in silicon.

2.7.1 Diffusion of carriers

Let us suppose that a concentration gradient of excess minority carriers is created in a rod of n-type semiconductor by injecting holes into one end of the rod via a suitable contact, as shown in Fig. 2.18. Owing to the random thermal motion of the holes at any given location it is probable that more holes will move to the right than to the left in a given time interval, and there will be a net movement or *diffusion* of holes along the rod down the concentration gradient. The net rate of flow of holes across unit area due to diffusion is found to be proportional to the concentration gradient, that is

$$\text{hole flux} = -D_h \frac{dp(x)}{dx} \quad (2.39)$$

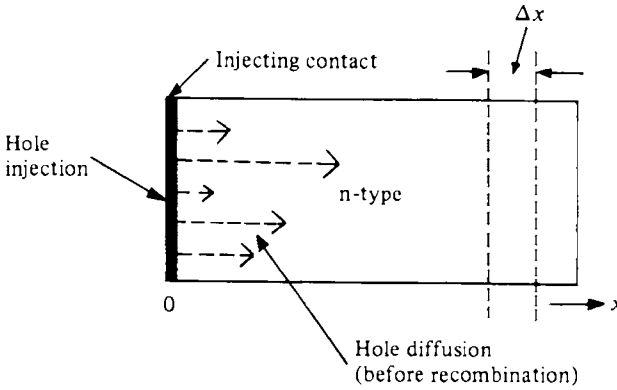


FIG. 2.18 Minority carrier injection and diffusion.

Similarly for electrons we have

$$\text{electron flux} = -D_e \frac{dn(x)}{dx} \quad (2.39a)$$

D_h and D_e are the hole and electron *diffusion coefficients*. These parameters are related to the mobilities in that they are a measure of the ease of carrier motion through the crystal lattice. In fact the so-called Einstein relationships give the diffusion coefficients in terms of the mobilities as (ref. 2.10)

$$D_{e,h} = \mu_{e,h} \frac{kT}{e} \quad (2.40)$$

As the holes diffuse along the rod they will eventually recombine. We now show that before recombining, the holes travel a characteristic distance called the *diffusion length* L_h . Let us consider a length Δx of the rod situated a distance x from the injecting contact (as shown in Fig. 2.18); then if the cross-sectional area of the rod is A the flow rate of holes into the left-hand face of the element is

$$-D_h \left(\frac{d}{dx} \Delta p(x) \right) A$$

Similarly, the flow rate out of the right-hand face is

$$-D_h \left[\left(\frac{d}{dx} \Delta p(x) \right)_x + \frac{d}{dx} \left(\frac{d}{dx} \Delta p(x) \right)_x \Delta x \right] A$$

The difference of these two terms is the net rate at which the element gains holes, which in the steady state must equal the rate of recombination of excess holes; therefore,

$$-D_h \frac{d^2(\Delta p(x))}{dx^2} \Delta x A = -\frac{\Delta p(x) \Delta x A}{\tau_h}$$

or

$$\frac{d^2 \Delta p(x)}{dx^2} - \frac{\Delta p(x)}{\tau_h D_h} = 0 \quad (2.41)$$

The solution to this equation, which is often called the steady state diffusion equation, is

$$\Delta p(x) = B_1 \exp(x/L_h) + B_2 \exp(-x/L_h)$$

where $L_h = \sqrt{D_h \tau_h}$ is the hole *diffusion length*. The constants B_1 and B_2 are determined by the boundary conditions, which in the case we are describing we can express as

$\Delta p(x) \rightarrow 0$ as $x \rightarrow \infty$, which implies that B_1 must be zero; and

$\Delta p(x) = \Delta p(0)$ at $x = 0$, so that $B_2 = \Delta p(0)$

Hence we have

$$\Delta p(x) = \Delta p(0) \exp(-x/L_h) \quad (2.42)$$

Thus we see that the excess minority carrier concentration decreases exponentially with distance; we shall return to this relationship in section 2.8.2.

2.7.2 Diffusion and drift of carriers

The diffusion of charge carriers will obviously give rise to a current flow. From eqs (2.39) we can write the electron and hole diffusion current densities as

$$J_e(\text{diff}) = e D_e \frac{dn}{dx} \quad (2.43)$$

$$J_h(\text{diff}) = -e D_h \frac{dp}{dx} \quad (2.43a)$$

If an electric field is present in addition to a concentration gradient, we can use eqs (2.24) and (2.43) to write the current densities for electrons and holes as

$$J_e = e \mu_e n \mathcal{E} + e D_e \frac{dn}{dx}$$

$$J_h = e \mu_h p \mathcal{E} - e D_h \frac{dp}{dx}$$

and the *total* current density is the sum of these contributions, that is

$$J = J_e + J_h$$

2.8

Junctions

The majority of the most useful electronic devices contain junctions between dissimilar materials, which may be metal-metal, metal-semiconductor or semiconductor-semiconductor

combinations. We shall concentrate our attention initially on *p-n homojunctions*, in which a junction is formed between p and n variants of the *same* semiconductor. Such junctions, in addition to their rectifying properties, form the basis of photodiodes, light-emitting diodes and photovoltaic devices, which will be discussed in later chapters. Many other devices, for example bipolar transistors and thyristors, contain two or more such junctions.

2.8.1 The p-n junction in equilibrium

A p-n junction may be fabricated as a single crystal of semiconductor by a number of different techniques (ref. 2.11). Indeed the exact behaviour of a junction depends to a large extent on the fabrication process used, which in turn determines the distances over which the change from p- to n-type nature occurs. For mathematical convenience we shall assume that the junction is *abrupt*, that is there is a step change in impurity type as shown in Fig. 2.19, which also shows the corresponding carrier concentrations p_p , n_p , and n_n , p_n on the p and n sides

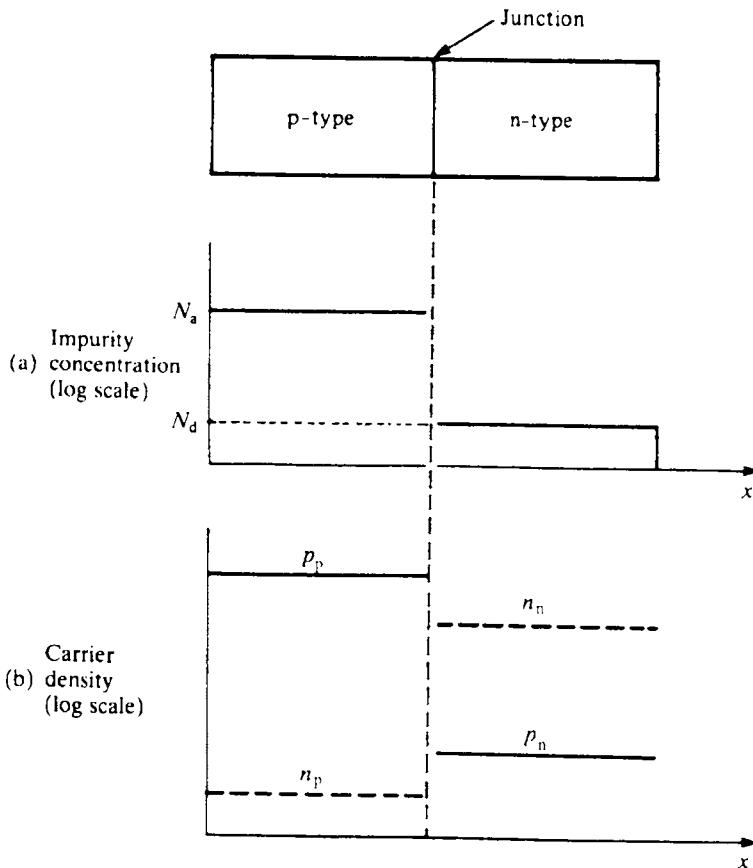


FIG. 2.19 Abrupt p-n junction showing (a) the impurity concentrations and (b) the carrier densities on the two sides of the junction.

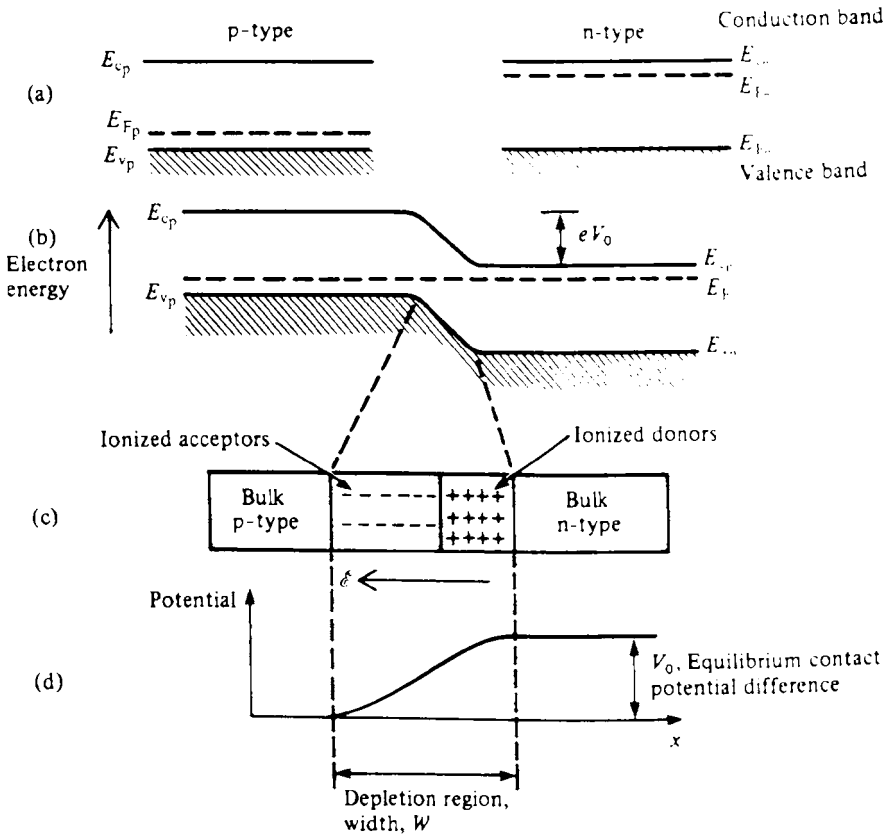


FIG. 2.20 Schematic representation of the formation of a p-n junction. (a) initially separated p-type and n-type materials; (b) the energy band distribution after the junction is formed; (c) the space charge layers of ionized impurity atoms within the depletion region W ; and (d) the potential distribution at the junction.

of the junction respectively. These apply only at relatively large distances from the junction; close to the junction they are modified as we shall see below.

While the assumption of a step junction may be used to establish many of the characteristics of junctions, it is not universally applicable. For example, in many cases the junction approximates to a *linearly graded* one in which there is a gradual change in doping type. One or two of the ways in which such junctions differ from abrupt ones will be pointed out in due course.

Although not the case in practice, let us assume that the junction is formed by bringing initially isolated pieces of n-type and p-type materials into intimate contact. Then, since there are many more holes in the p-type than in the n-type material, holes will diffuse from the p to the n region. The holes diffusing out of the p-type side leave behind 'uncovered' or ionized acceptors, thereby building up a negative space charge layer in the p-type side close to the junction. Similarly electrons, diffusing into the p-type side, leave behind a positive space charge layer of ionized donors as shown in Fig. 2.20(c). This double space charge layer causes

an electric field to be set up across a narrow region on either side of the junction, directed from the n- to the p-type region as shown.

The direction of the junction electric field is such as to inhibit further diffusion of the majority carriers, though such diffusion is not prevented altogether. This must be so since the electric field will sweep minority carriers across the junction so that there is a drift current of electrons from the p- to the n-type side, and of holes from the n- to the p-type side, which is in the opposite direction to the diffusion current. The junction field thus builds up until the drift and diffusion current flows are equal in magnitude, at which stage the Fermi level is constant across the junction as shown in Fig. 2.20(b) indicating that equilibrium has been reached within the crystal as a whole. Thus, as there is no *net* current flow in equilibrium,

$$J_h(\text{drift}) + J_h(\text{diff}) = 0$$

and

$$J_e(\text{drift}) + J_e(\text{diff}) = 0$$

The induced electric field establishes a *contact* or *diffusion* potential V_0 between the two regions and the energy bands of the p-type side are displaced relative to those of the n-type as shown in Fig. 2.20(b). The magnitude of the contact potential depends on the temperature and the doping levels as we shall see below. The contact potential is established across the space charge region, which is also referred to as the *transition* or *depletion* region, so called because this region has been almost depleted of its majority carriers. As a consequence it is very resistive relative to the other (so-called bulk) regions of the device.

An expression relating the contact potential to the doping levels can be obtained, for example, from eq. (2.33). From this equation and adopting the notation used in Figs 2.19 and 2.20 we can write the electron concentration in the conduction band of the p-type side as

$$n_p = N_c \exp \left[- \left(\frac{E_{c_p} - E_{F_p}}{kT} \right) \right]$$

Similarly the electron concentration in the n-type side is

$$n_n = N_c \exp \left[- \left(\frac{E_{c_n} - E_{F_n}}{kT} \right) \right]$$

As we mentioned above, the Fermi level is constant everywhere in equilibrium so that $E_{F_p} = E_{F_n} = E_F$ say. Hence, eliminating N_c we have

$$E_{c_p} - E_{c_n} = kT \ln \left(\frac{n_n}{n_p} \right) = eV_0$$

Therefore

$$V_0 = \frac{kT}{e} \ln \left(\frac{n_n}{n_p} \right) \quad (2.44)$$

At temperatures in the range $100 \text{ K} \leq T \leq 400 \text{ K}$ the majority carrier concentrations are equal to the doping levels, that is $n_n = N_d$ and $p_p = N_a$, and remembering $np = n_i^2$ we can write eq. (2.44) as

$$V_0 = \frac{kT}{e} \ln \left(\frac{N_a N_d}{n_i^2} \right) \quad (2.45)$$

EXAMPLE 2.5 Equilibrium p-n junction contact potential difference

The contact potential difference V_0 in a given germanium diode may be calculated from the following data: donor impurity level $N_d = 10^{22} \text{ m}^{-3}$; acceptor impurity level $N_a = 10^{24} \text{ m}^{-3}$; and intrinsic electron concentration $n_i = 2.4 \times 10^{19} \text{ m}^{-3}$.

Then assuming room temperature ($T \approx 290 \text{ K}$) we have, from eq. (2.45), that

$$V_0 = 0.025 \ln \left(\frac{10^{22} \times 10^{24}}{(2.4 \times 10^{19})^2} \right) = 0.42 \text{ V}$$

Equation (2.44) gives us a very useful relationship between the carrier concentrations on the two sides of the junction, that is

$$n_p = n_n \exp \left(\frac{-eV_0}{kT} \right) \quad (2.46)$$

and similarly

$$p_n = p_p \exp \left(\frac{-eV_0}{kT} \right) \quad (2.46a)$$

2.8.2 Current flow in a forward-biased p-n junction

If the equilibrium situation is disturbed by connecting a voltage source externally across the junction there will be a net current flow. The junction is said to be *forward biased* if the p region is connected to the positive terminal of the voltage source as shown in Fig. 2.21(a). As we mentioned in the last section, the depletion region is very resistive in comparison with the bulk regions so that the external voltage V is dropped almost entirely across the depletion region. This has the effect of lowering the height of the potential barrier to $(V_0 - V)$ as shown in Fig. 2.21(b). Consequently, majority carriers are able to surmount the potential barrier much more easily than in the equilibrium case so that the diffusion current becomes much larger than the drift current. There is now a net current from the p to the n region in the conventional forward sense and carriers flow in from the external circuit to restore equilibrium in the bulk regions. We note that with the application of an external potential the Fermi levels are no longer aligned across the junction.

The reduction in height of the potential barrier leads to majority carriers being *injected*

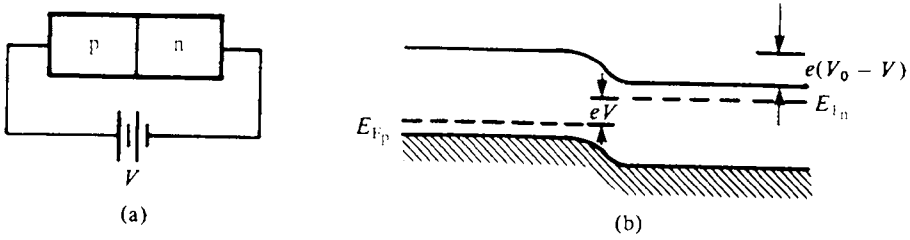


FIG. 2.21 Forward bias voltage V applied to a p-n junction (a) and the resulting energy band structure (b).

across the junction. On being so injected, these carriers immediately become minority carriers, and the minority carrier concentrations near to the junction rise to the new values n'_p and p'_n . This establishes excess minority carrier concentration gradients, as shown in Fig. 2.22, so that the injected carriers diffuse away from the junction. This situation is precisely the same as that described in section 2.7.1 and thus, considering the n region, the injected holes diffuse away from the junction recombining as they do so. The electrons lost in this recombination are replaced by the external voltage source so that a current flows in the external circuit. A similar argument applies to the p region, with the roles of electrons and holes reversed. It should be noted that the majority carrier concentrations are not noticeably changed as a consequence of the injection (Fig. 2.22) unless the bias voltage is almost equal to V_0 , resulting in a very large current flow.

The drift current is relatively insensitive to the height of the potential barrier since all the minority carriers generated within about a diffusion length of the edge of the depletion region may diffuse to the depletion region and be swept across it, whatever the size of the electric field there.

By the same argument that was used to give eqs (2.46), we see that with forward bias the minority carrier concentrations in the bulk regions adjacent to the depletion layer

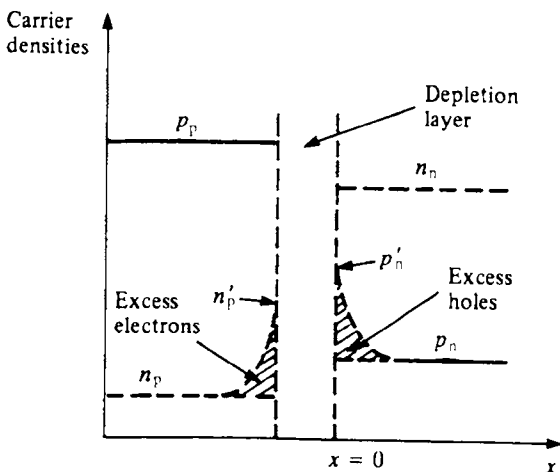


FIG. 2.22 Carrier densities in the bulk region of a forward-biased p-n junction diode. Owing to carrier injection, the minority carrier densities close to the depletion layer are *greater* than the equilibrium values.

become

$$n'_p = n_n \exp\left(\frac{-e(V_0 - V)}{kT}\right) \quad (2.47)$$

and

$$p'_n = p_p \exp\left(\frac{-e(V_0 - V)}{kT}\right) \quad (2.47a)$$

Then taking eqs (2.46a) and (2.47a), for example, we have

$$p'_n = p_n \exp\left(\frac{eV}{kT}\right) \quad (2.48)$$

As we noted above, the excess minority carrier concentration will decrease owing to recombination in accordance with eq. (2.42) so that we may write the excess hole concentration in the n region at a distance x from the edge of the depletion layer as

$$\Delta p(x) = \Delta p(0) \exp(-x/L_h)$$

where now

$$\Delta p(x) = p'_n(x) - p_n$$

and $\Delta p(0)$ is the value of $\Delta p(x)$ at $x = 0$, that is $(p'_n - p_n)$. Using eq. (2.48) we can therefore write

$$\Delta p(0) = p_n [\exp(eV/kT) - 1] \quad (2.49)$$

We have assumed a one-dimensional carrier flow in the x direction, which is an acceptable approximation even though in practice carrier flow occurs in three dimensions.

Now, we have argued that the electric fields in the bulk regions are very small and therefore, adjacent to the depletion layer in the n region, and in particular at $x = 0$, the total current density will be due to diffusion only. Thus the current density due to hole motion is given by eq. (2.43b). Hence differentiating eq. (2.42) and substituting into eq. (2.43b) gives

$$J_h = \frac{eD_h}{L_h} \Delta p(0) \exp\left(-\frac{x}{L_h}\right)$$

At $x = 0$, we can write this, using eq. (2.49), as

$$J_h = \frac{eD_h}{L_h} p_n \left[\exp\left(\frac{eV}{kT}\right) - 1 \right] \quad (2.50)$$

There is a similar contribution due to electron flow, that is

$$J_e = \frac{eD_e}{L_e} n_p \left[\exp\left(\frac{eV}{kT}\right) - 1 \right] \quad (2.50a)$$

The total current density is therefore given by

$$J = J_0 \left[\exp\left(\frac{eV}{kT}\right) - 1 \right] \quad (2.51)$$

where

$$J_0 = e \left(\frac{D_h}{L_h} p_n + \frac{D_e}{L_e} n_p \right) \quad (2.51a)$$

J_0 is called the saturation current density; it represents the current flow when a few volts of reverse bias is applied to the junction (see section 2.8.3).

EXAMPLE 2.6 Saturation current density

We may calculate the saturation current density in an abrupt silicon junction given the following data: $N_d = 10^{21} \text{ m}^{-3}$; $N_a = 10^{22} \text{ m}^{-3}$; $D_e = 3.4 \times 10^{-3} \text{ m}^2 \text{ s}^{-1}$; $D_h = 1.2 \times 10^{-3} \text{ m}^2 \text{ s}^{-1}$; $L_e = 7.1 \times 10^{-4} \text{ m}$; $L_h = 3.5 \times 10^{-4} \text{ m}$; and $n_i = 1.6 \times 10^{16} \text{ m}^{-3}$.

Assuming that all of the impurities are ionized, we have $n_n = N_d = 10^{21} \text{ m}^{-3}$ and therefore

$$p_n = \frac{2.56 \times 10^{32}}{10^{21}} = 2.56 \times 10^{11} \text{ m}^{-3}$$

Similarly $n_p = 2.56 \times 10^{10} \text{ m}^{-3}$. Therefore from eq. (2.51a) $J_0 = 1.6 \times 10^{-7} \text{ A m}^{-2}$.

A typical discrete diode may have a junction area of about 10^{-6} m^2 and hence the reverse bias saturation current would be $i_0 = 1.6 \times 10^{-13} \text{ A}$.

2.8.3 Current flow in a reverse-biased p-n junction

In this case the external bias is applied so that the p region is connected to the negative terminal of the voltage source as shown in Fig. 2.23(a). This has the effect of increasing the

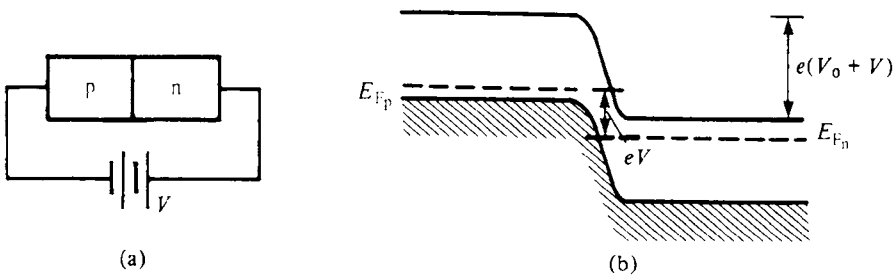


FIG. 2.23 Reverse bias voltage V applied to a p-n junction (a) and the resulting energy band structure (b).

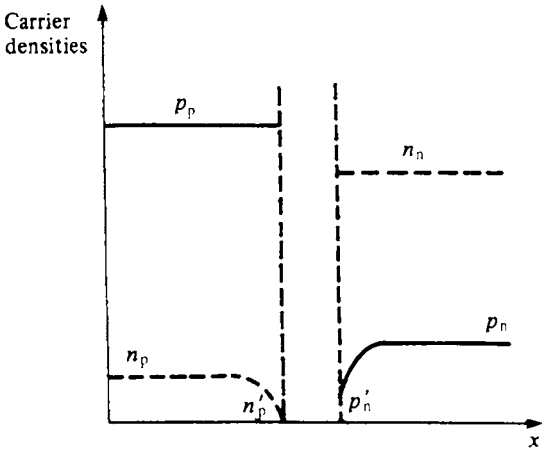


FIG. 2.24 Carrier densities in the bulk regions of a reverse-biased p-n junction diode. Owing to carrier extraction the minority carrier densities close to the depletion layer are less than the equilibrium values.

height of the potential barrier to $V_0 + V$ (Fig. 2.23b), thereby reducing the diffusion current to negligible proportions. The net current flow is therefore the drift current which is directed in the conventional reverse sense, that is from the n to the p region. This results in carrier extraction rather than injection because the minority carriers generated near the junction diffuse to it and are swept across the depletion region. The nearer the carriers are generated to the junction the greater is the probability of this occurring so that a concentration gradi-

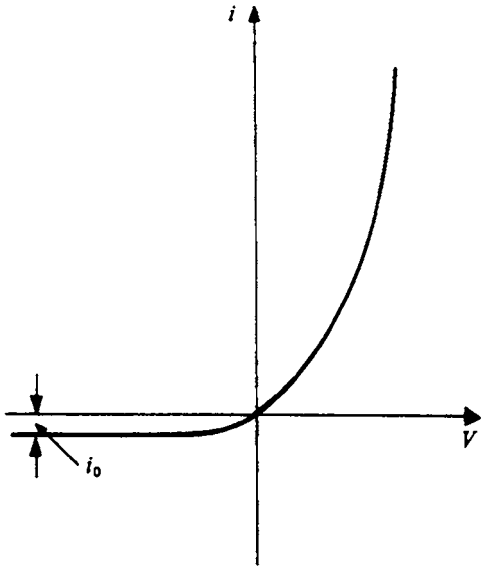


FIG. 2.25 Current-voltage characteristics of a p-n junction diode. The reverse saturation current i_0 is equal to J_0 multiplied by the junction cross-sectional area.

ent is formed towards the junction, as illustrated in Fig. 2.24. The drift of carriers across the junction is therefore 'fed' by diffusion so that we may use precisely the same arguments to derive the current – voltage relationship as we used in the previous section. Equations (2.51) also apply in this case, the only difference being, of course, that the sign of V is changed.

The current – voltage characteristic of an ideal p–n junction is therefore as shown in Fig. 2.25; in forward bias the current increases exponentially with voltage, while in reverse bias the current saturates at J_0 times the junction cross-sectional area.

2.8.4 Junction geometry and depletion layer capacitance

The two space charge layers at the junction vary in width and therefore in the amount of charge they contain as the bias voltage changes, thereby giving rise to an effective depletion layer (or junction) capacitance $C_j = dQ_j/dV$. This capacitance together with a charge storage capacitance (see below) is one of the factors which limits the high frequency operation of junction devices. On the other hand, the capacitance can be controlled by the bias voltage, and this effect is exploited in voltage-dependent capacitors such as the varactor diode.

The charge density within the space charge layers results simply from the charge on the ionized impurities. Thus the charge density within the p region is

$$\rho_p = -N_a e \quad (2.52)$$

and similarly

$$\rho_n = N_d e \quad (2.52a)$$

Referring to Fig. 2.26, we denote the widths of the p and n space charge regions by x_p and x_n , with the origin of x at the actual junction. Using Poisson's equation we may calculate the electric field distribution within the depletion layer. For the p side, from eq. (2.52), we have

$$\frac{d\mathcal{E}}{dx} = -\frac{eN_a}{\epsilon_0 \epsilon_r} \quad -x_p < x < 0 \quad (2.53)$$

where ϵ_r is the relative permittivity of the semiconductor. Integrating eq. (2.53) with the appropriate limits, that is

$$\int_0^{x_p} d\mathcal{E} = -\frac{eN_a}{\epsilon_0 \epsilon_r} \int_{-x_p}^0 dx$$

gives \mathcal{E}_0 , the maximum value of the junction electric field. We find that

$$\mathcal{E}_0 = \frac{-eN_a x_p}{\epsilon_0 \epsilon_r} \quad (2.54)$$

Similarly for the n side we have

$$\mathcal{E}_0 = \frac{-eN_d x_n}{\epsilon_0 \epsilon_r} \quad (2.54a)$$

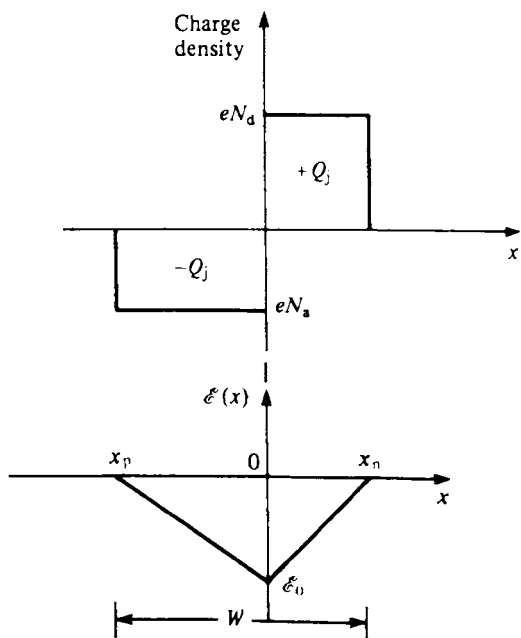


FIG. 2.26 Space charge density and variation of electric field \mathcal{E} within the depletion region of a p-n junction. Note that as $N_d > N_a$ in this example, then $x_n < x_p$, but the space charge on each side of the junction has the same magnitude.

From eqs (2.54) we have

$$N_a x_p = N_d x_n \tag{2.55}$$

from which we see that the depletion layer extends farthest into the least heavily doped side of the junction as we have implied in, for example, Figs 2.20 and 2.26. This is a result of some significance in explaining the operation of many devices.

We can now relate the electric field to the junction contact potential V_0 . Since in general

$$\mathcal{E}(x) = - \frac{dV(x)}{dx}$$

we may write

$$- \int_0^{V_0} dV(x) = -V_0 = \int_{-x_p}^{x_n} \mathcal{E}(x) dx$$

The right-hand integral is the area of the $\mathcal{E}(x)$ versus x triangle shown in Fig. 2.26 so that

$$-V_0 = \frac{1}{2} W \mathcal{E}_0 \tag{2.56}$$

where the width of the depletion layer W is

$$W = |x_p| + |x_n| \tag{2.57}$$

Substituting for ϵ_0 from eq. (2.54a) into eq. (2.56) gives

$$V_0 = \frac{eWN_d x_n}{2\epsilon_0 \epsilon_r} \quad (2.58)$$

while combining eqs (2.55) and (2.57) gives

$$x_n = \frac{WN_a}{N_d + N_a} \quad (2.59)$$

Then substituting x_n from eq. (2.59) into eq. (2.58) yields

$$V_0 = \frac{e}{2\epsilon_0 \epsilon_r} \left(\frac{N_a N_d}{N_d + N_a} \right) W^2$$

If a bias voltage is applied, the above argument still holds but we must replace V_0 by $(V_0 - V)$, where V is positive for forward bias and negative for reverse bias. Therefore, finally, we have

$$W = \left[\frac{2\epsilon_0 \epsilon_r}{e} (V_0 - V) \left(\frac{1}{N_a} + \frac{1}{N_d} \right) \right]^{1/2} \quad (2.60)$$

The capacitance associated with the depletion layer can be obtained as follows. The stored charge Q_j on either side of the junction is given by

$$|Q_j| = AeN_d x_n = AeN_a x_p \quad (2.61)$$

where A is the junction area. Hence using eqs (2.59) and (2.60) we may write

$$|Q_j| = \frac{AeN_d N_a W}{N_d + N_a} = A \left[2\epsilon_0 \epsilon_r e (V_0 - V) \left(\frac{N_d N_a}{N_d + N_a} \right) \right]^{1/2} \quad (2.62)$$

If the bias voltage V changes then Q_j changes so that the junction capacitance C_j is given by differentiating eq. (2.62) with respect to V . Hence

$$C_j = \frac{dQ_j}{dV} = \frac{A}{2} \left[\frac{2e\epsilon_0 \epsilon_r}{(V_0 - V)} \left(\frac{N_d N_a}{N_d + N_a} \right) \right]^{1/2} \quad (2.63)$$

It is left as an exercise for the reader to show that $C_j = A\epsilon_0 \epsilon_r / W$ farads, indicating that the abrupt junction behaves rather like a parallel plate capacitor of plate separation W and area A containing a medium of relative permittivity ϵ_r . We note, for reasonably large values of reverse bias V , that $C \propto V^{-1/2}$. In contrast, for the case of a graded junction it is found that $C \propto V^{-1/3}$.

In forward bias the junction capacitance C_j is swamped by another capacitive effect which gives rise to the charge storage or diffusion capacitance C_d . In forward bias, many carriers are injected across the junction and it is the time taken for this injected carrier density to adjust to changes in the forward bias voltage which gives rise to the charge storage capacitance. The principal mechanism whereby the injected carrier density adjusts is recombination, and consequently the expression for C_d includes the carrier lifetime τ , that is

$$C_d = eI\tau/kT \text{ (see ref. 2.12).}$$

EXAMPLE 2.7 Junction depletion layer capacitance

We may calculate the capacitance of a silicon $p^+ - n$ junction at a reverse bias of 4 V, given that $N_d = 4 \times 10^{21} \text{ m}^{-3}$, $V_0 = 0.8 \text{ V}$, $\epsilon_r = 11.8$ and that the junction area is $4 \times 10^{-7} \text{ m}^2$.

In a $p^+ - n$ junction, $N_a \gg N_d$ and hence eq. (2.63) reduces to

$$C_j = \frac{A}{2} \left[\frac{2e\epsilon_0\epsilon_r N_d}{V_0 - V} \right]^{1/2}$$

With reverse bias $V = -4 \text{ V}$ we find that

$$C_j = 33.4 \text{ pF}$$

2.8.5 Deviations from simple theory

Although the theory described above and the diode characteristic shown in Fig. 2.25 are in reasonable agreement with what is observed in practice, there are several points of difference. For our purposes one of the most important deviations is that, at sufficiently large values of reverse bias, *breakdown* occurs. That is, there is a sudden and rapid increase in reverse current at a particular value of reverse bias voltage (e.g. Fig. 2.27b).

Reverse breakdown occurs by two mechanisms. The first, called the *Zener effect*, is due to quantum mechanical tunnelling. This takes place most readily in heavily doped junctions, which from eq. (2.60) leads to narrow depletion layers and therefore high junction fields.

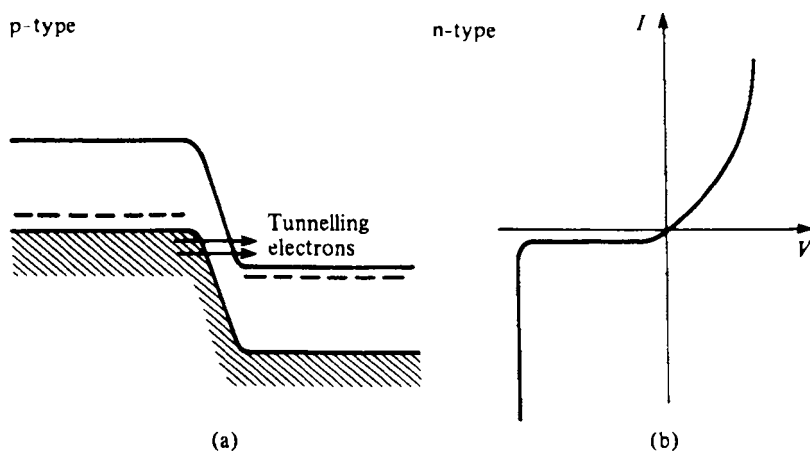


FIG. 2.27 Schematic representation of electron tunnelling in Zener breakdown of a heavily doped $p - n$ junction under reverse bias (a) and the resulting effect on the $i - V$ characteristic (b).

In effect, as we can see from Fig. 2.27(a), the energy bands on the two sides of the junction become 'crossed' so that filled states in the valence band of the p side are aligned with empty states in the conduction band of the n side. Electrons, therefore, tunnel from the p to the n side vastly increasing the reverse current.

The second mechanism, *avalanche breakdown*, occurs in lightly doped junctions with wide depletion layers. This mechanism involves impact ionization of the host atoms by energetic carriers. If carriers crossing the depletion layer acquire sufficient energy from the electric field between collisions, they may ionize lattice atoms on colliding with them. The electrons and holes so produced may, in turn, cause further ionizing collisions and so on to generate an avalanche of carriers. Neither breakdown mechanism is in itself destructive to the junction. If, however, the reverse current is allowed to become too large then Joule heating may cause damage to the device.

Other important deviations arise from our having ignored such factors as carrier generation and recombination within the depletion layer. Carrier recombination leads to an increase in current in the forward direction as the carriers which recombine must be replaced from the external circuit. In silicon and gallium arsenide diodes, this current mechanism may be more important than current flow due to carrier injection, especially at low currents. This and other factors lead in some cases to the p–n junction having a current–voltage relationship of the form $J = J_0[\exp(eV/\beta kT) - 1]$, where β varies between 1 and 2 depending on the semiconductor and the temperature. Because β determines the departure from the ideal diode characteristic, it is often called the *ideality factor*.

Carrier generation in the depletion layer gives rise to a larger value of reverse bias current than the simple theory predicts. Optical generation of carriers within the depletion layer may give rise to an increase in reverse current or initiate avalanche breakdown if the reverse bias is sufficiently great. These phenomena form the bases of the photodiodes discussed in Chapter 7.

2.8.6 Other junctions

In the discussion of p–n junctions in the preceding sections the basic material has been the same on both sides of the junction, that is we have been discussing the behaviour of semiconductor homojunctions. In many devices, however, this is not the case and heterojunctions are used with significant advantage. Thus, for example, we have semiconductor heterojunctions and metal–semiconductor junctions. We shall briefly describe the basic theory of both of these types of junction. It is interesting to note that in both cases the predictions of the basic theory are not always borne out in practice; this is often caused by the behaviour of defects close to the junction (ref 2.13).

2.8.6.1 Semiconductor heterojunctions

Semiconductor heterojunctions are widely used in LEDs and particularly in semiconductor lasers (see section 5.10.2). Such junctions are formed by growing a particular semiconductor crystal onto another *different* semiconductor crystal. This is relatively easy providing that the crystals both have the same crystal structure with comparable lattice parameters. At the

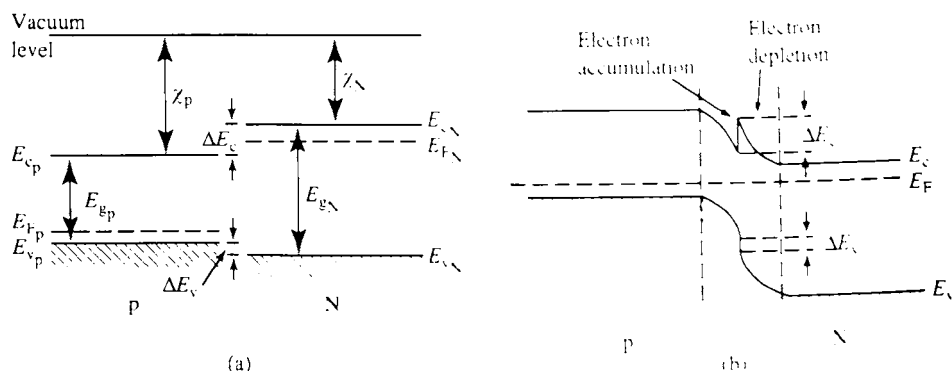


FIG. 2.28 A p-N heterojunction: (a) shows the energy bands of the semiconductors separately; (b) shows the energy bands after junction formation. (For a GaAs/Ga_{0.45}Al_{0.55}As heterojunction, $E_g = 1.43$ eV, $E_{gN} = 1.8$ eV, $\Delta E_c = 0.32$ eV and $\Delta E_v = 0.05$ eV.)

boundary between the two materials, however, there is a change in such material parameters as the energy gap E_g , the work function ϕ , the electron affinity χ and the relative permittivity ϵ_r (and hence also the refractive index n).

At a heterojunction each of the semiconductors may be doped n-type or p-type so there are four possible combinations, namely n-N, p-P (*isotype*) and n-P, P-n (*unisotype*) junctions, where the capital letter denotes the wider bandgap material. Several devices use heterojunctions formed between gallium arsenide and gallium aluminium arsenide, Ga_xAl_{1-x}As, and we shall base our discussion on these materials. It should be noted that the values of the material parameters mentioned above for gallium aluminium arsenide depend on the precise composition, that is on the value of x .

Let us first consider a p-N heterojunction. Figure 2.28(a) defines the characteristics of the materials separately. Figure 2.28(b) illustrates the electron energy bands of the junction in equilibrium, which at first glance may seem to be rather confusing. If it is recalled, however, that mobile charges flow across the junction, leaving behind ionized impurities, until equilibrium is reached (i.e. the Fermi level is constant across the junction), and that the energy gap of a semiconductor remains constant within the semiconductor for constant doping, then the situation becomes clearer. Electrons have flowed from regions of high to low potential so that a depletion layer is formed in the conduction band of the N-type gallium aluminium arsenide and the bands bend upwards. Correspondingly an electron accumulation layer has formed in the p-type gallium arsenide so that the band bends downwards. An energy spike of height ΔE_c is formed between these regions; the precise value of ΔE_c is determined by the values of E_g and χ for the two materials on either side of the junction. Similarly a discontinuity of height ΔE_v appears in the valence bands at the junction. We would expect such junctions to be rectifying, and indeed they are in practice. Similarly P-n junctions, as illustrated in Fig. 2.29, are expected to be rectifying, and indeed are found to be so in practice. The same is not always the case with n-N and p-P junctions based on these materials, which frequently exhibit ohmic characteristics despite expectations to the contrary.

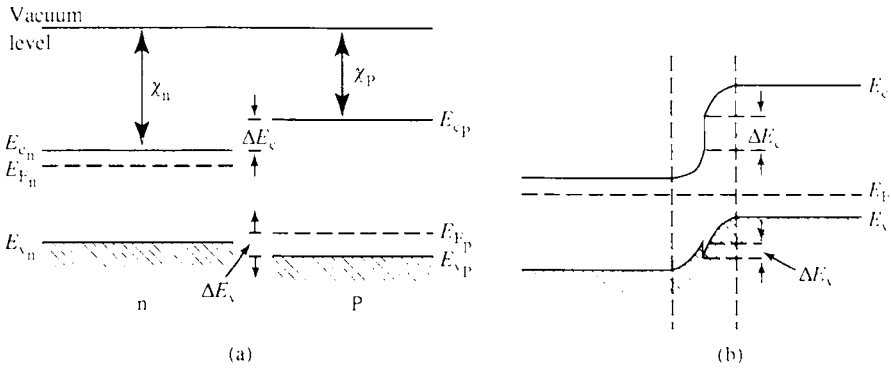


FIG. 2.29 An n-p heterojunction before (a) and after (b) junction formation.

The potential barriers created at heterojunctions can be significantly larger than in the case of homojunctions; thus, for example, electrons injected into the narrow bandgap material (gallium arsenide) tend to accumulate there. Consequently, as we shall see in section 5.10.2, if an appropriate double heterojunction structure is formed carriers can be confined in the region between the two heterojunctions and thus form the active region of a semiconductor laser. A further rather obvious, but useful, factor is that radiation created by band-to-band electron-hole recombinations in the wide bandgap material will not be absorbed by the narrow bandgap material. This is important in the fabrication of efficient surface- and edge-emitting LEDs. Further details of these and related benefits resulting from the use of heterojunctions are described in ref. 2.13.

2.8.6.2 Metal-semiconductor junctions

Metal-semiconductor junctions are attractive because of their ease of fabrication, and, as we shall see in section 7.3.6.5, they are useful when operation with a fast response is required. The behaviour of ideal metal-semiconductor junctions depends largely on the doping type of the semiconductor and the relative size of the work functions of the two materials.

Let us first consider a metal to n-type semiconductor junction in which the work function of the metal is greater than that of the semiconductor (i.e. $\phi_m > \phi_s$). The energy bands of the separated materials are illustrated in Fig. 2.30(a), while the situation after the junction is formed is shown in Fig. 2.30(b). As usual, when contact is established, charge flows across the junction until the Fermi levels are aligned in equilibrium. In this case electrons flow from the semiconductor into the metal thereby creating a depletion layer in the semiconductor close to the junction. The accumulated negative charge in the metal is balanced by the positive charge on the uncompensated donor ions in the semiconductor. The resulting junction electric field, contact potential difference and band bending are similar to the effects already discussed for p-n junctions in section 2.8.1. The depletion layer width, for example, may be calculated from eq. (2.60) by treating the junction as though it were a p⁺-n junction.

The equilibrium potential difference V_0 , which prevents further net electron diffusion across

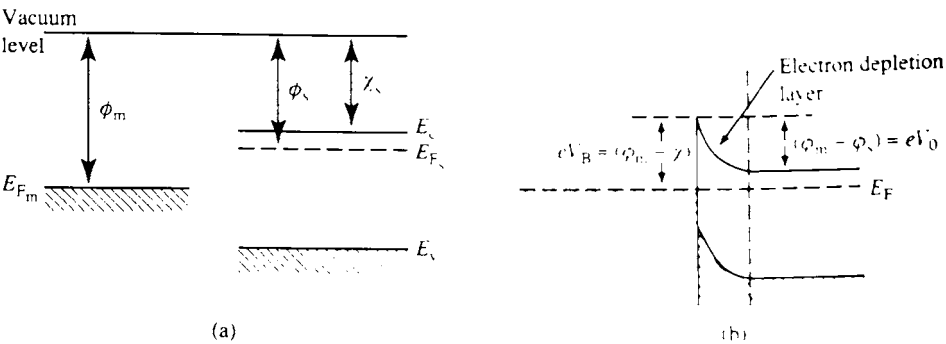


FIG. 2.30 A Schottky barrier formed by contacting a metal to an n-type semiconductor with the metal having the larger work function: band diagrams (a) before and (b) after the junction is formed.

the junction from the n-type semiconductor to the metal, is given by $(\phi_m - \phi_s)$ as shown in Fig. 2.30(b). The potential barrier height V_B for electron injection from the metal into the semiconductor is slightly different, and is given by $\phi_m - \chi_s$, where χ_s is the electron affinity of the semiconductor. The potential barrier V_0 can be decreased or increased by the application of an external forward or reverse bias voltage respectively. In either case V_B remains constant so that the junction behaves as a rectifying one as illustrated in Fig. 2.31. The junction between a metal and p-type semiconductor, where $\phi_m < \phi_s$, is also rectifying; both of these rectifying junctions are referred to as Schottky barrier diodes.

In both cases the forward current is due to the injection of *majority* carriers from the semiconductor into the metal; this is in contrast to the forward current flow in p-n junctions. The absence of minority carrier injection and the associated charge storage delay time (which is determined by the carrier recombination time) results in a very much reduced capacitance, which is important, for example, in the behaviour of fast Schottky barrier photodiodes with short response times.

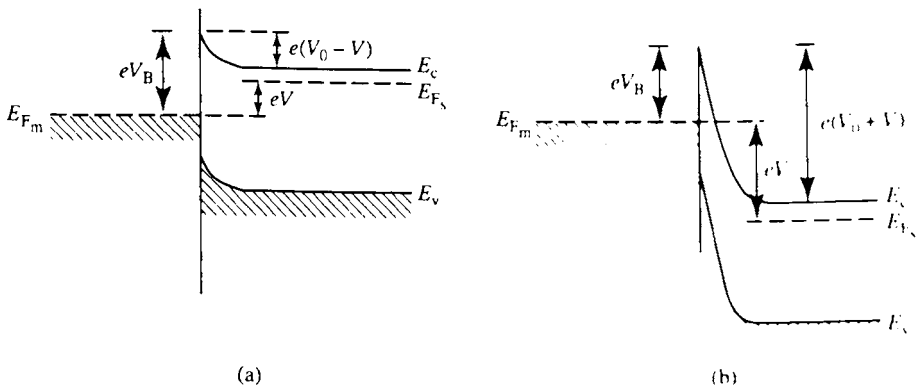


FIG. 2.31 The metal to n-type semiconductor junction shown in Fig. 2.30 under (a) forward bias and (b) reverse bias.

In contrast let us consider the behaviour of metal to n-type semiconductor junctions in which $\phi_m < \phi_s$. The Fermi levels now become aligned as a consequence of electron flow from the metal into the semiconductor, thereby raising the semiconductor electron energies relative to those in the metal in equilibrium. In this case there is no depletion layer in the semiconductor, and there is only a small potential barrier opposing further electron flow from the metal to the semiconductor, which is easily overcome by the application of a small external bias. Thus in effect we have an ohmic junction in which the current flow is essentially independent of the direction of the bias voltage. The same is the case for junctions involving p-type semiconductors in which $\phi_m < \phi_s$.

While it is imperative that we should be able to form good ohmic contacts between semiconductor devices and external circuits, the ohmic contacts described above do not always behave as expected, and often display rectifying properties. This is largely because of the presence of *surface states* at the junction interface. Unlike semiconductor p-n junctions, which are formed within a single crystal, metal-semiconductor junctions include a termination of the semiconductor crystal. The surface of the semiconductor invariably includes surface states which can trap positive or negative charge at the surface, depending on the energies of the states. This trapped charge can cause band bending of the semiconductor energy bands resulting in the formation of a potential barrier (even in the absence of the metal) and in rectifying behaviour (ref. 2.14).

One way of ensuring good ohmic contacts is to create a very heavily doped p^+ or n^+ layer very close to the semiconductor surface. If a potential barrier exists at the surface the depletion layer then will be thin enough so that carriers can easily tunnel through it. Such heavy surface doping layers can be obtained by using appropriate metals which act as donors or acceptors as the contact material, for example aluminium on p-type silicon. Alternatively they can be created by including an additional diffusion or ion implementation stage in the fabrication process.

2.9 The quantum well

There has been considerable interest recently in devices based on semiconductor structures in the form of a very thin slab, in which the thickness is less than about 10 nm, and very much less than the length and breadth of the slab. Quantum well structures may, for example, be formed by sandwiching a region of gallium arsenide between two gallium aluminium arsenide regions; the gallium arsenide is then bounded by two heterojunctions. If electrons are confined within such a structure then the energy levels they occupy can be obtained from eq. (2.13), we suppose that the slab thickness is L_z , and that the other dimensions are L_x and L_y , so that

$$E(n_1, n_2, n_3) = \frac{h^2}{8m_e^*} \left(\frac{n_1^2}{L_x^2} + \frac{n_2^2}{L_y^2} + \frac{n_3^2}{L_z^2} \right) \quad (2.64)$$

where n_1, n_2, n_3 are positive integers greater than zero. There is of course a corresponding expression for holes in the valence band.

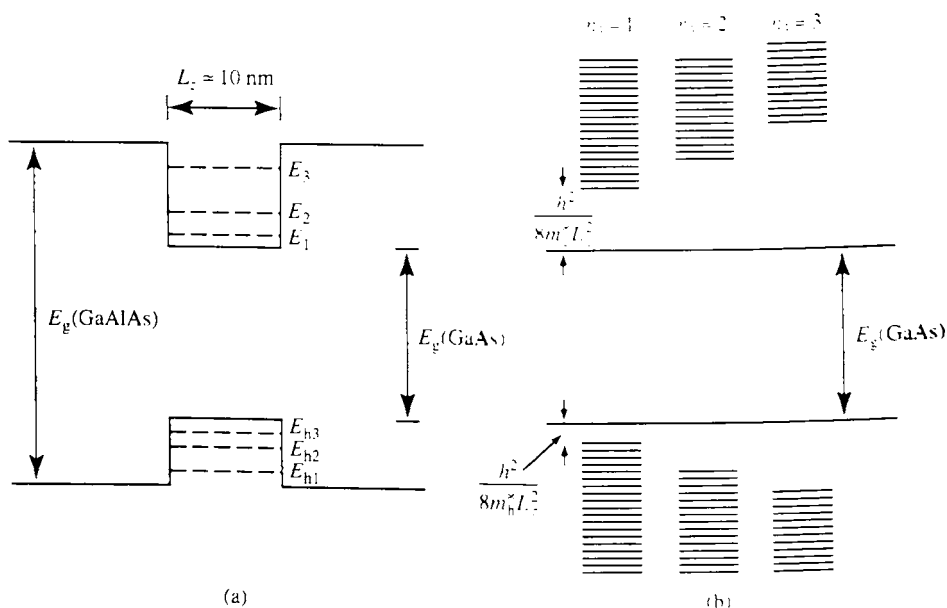


FIG. 2.32 Energy levels within a quantum well structure of GaAs sandwiched between two GaAlAs regions: (a) shows the discrete electron and hole energy levels for $n_1 = 1, 2, 3$ while (b) shows the subbands corresponding to these values of n_1 (they are separated for clarity, but in reality they are superimposed on one another).

Since L_z is so small compared with L_x and L_y , the energy levels will form groups, or subbands, of closely spaced levels determined by the various integral values of n_1 and n_2 , for each value of n_3 as shown in Fig. 2.32. The density of states distribution in this case is no longer the smooth parabolic curve depicted in Fig. 2.14(a), but rather it has the 'staircase' structure shown in Fig. 2.33. There are a number of consequences of this situation. Firstly we see that the lowest energy states, corresponding to $n_1 = 1$, are situated at an energy approximately $\hbar^2/8m_e^*L_z^2$ above the bottom of the conduction band, and correspondingly the uppermost states in the valence band are about $\hbar^2/8m_h^*L_z^2$ below the top of the band. Thus the effective energy gap between the bands has increased by an amount

$$\Delta E_g = \frac{\hbar^2}{8L_z^2} \left(\frac{1}{m_e^*} + \frac{1}{m_h^*} \right)$$

If we take $L_z = 10 \text{ nm}$, then as Example 2.8 shows, ΔE_g is of the order of 0.1 eV in GaAs.

EXAMPLE 2.8 Energy gap in a GaAs quantum well

We may calculate the effective increase in the width of the energy gap in a quantum well structure in GaAs, with a thickness L_z of 10 nm.

For GaAs

$$m_e^* = 0.068m \quad m_h^* = 0.56m$$

Hence

$$\begin{aligned} \Delta E_g &= \frac{(6.626 \times 10^{-34})^2}{8 \times (10 \times 10^{-9})^2 \times (9.1 \times 10^{-31})} \left(\frac{1}{0.068} + \frac{1}{0.56} \right) \\ &= 9.93 \times 10^{-21} \text{ J} \quad \text{or} \quad 0.062 \text{ eV} \end{aligned}$$

Secondly we see that there are more states towards the bottom of the conduction band than for the normal case (i.e. when L_x, L_y, L_z are all comparable).

Because of the relative magnitudes of L_x, L_y, L_z the lowest energy states will all have $n_3 = 1$, and will effectively constitute a 'two-dimensional' group of states with varying n_1 and n_2 values. Now while the density of states function for a normal three-dimensional solid is given by eq. (2.29) from which we see that $Z(E) \propto E^{1/2}$, the corresponding function for the two-dimensional case is given by $Z(E) = 4\pi m_e / h^2$, which is independent of energy (see Problem 2.13). The distribution of electrons over these states at temperature T is given by $n(E) = Z(E)F(E)$, which, assuming that $E - E_F \gg kT$, becomes

$$n(E) \approx \frac{4\pi m_e}{h^2} \exp \left[- \left(\frac{E - E_F}{kT} \right) \right] \quad (2.65)$$

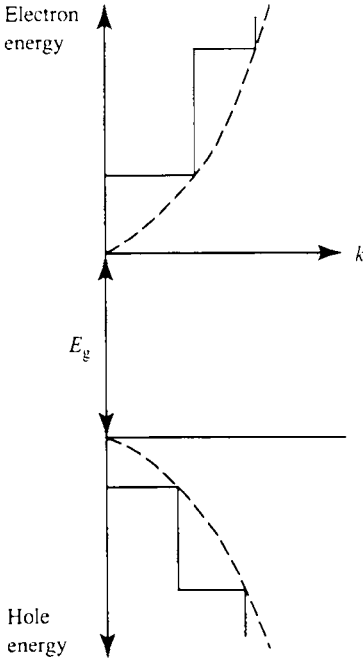


FIG. 2.33 The 'stairwell' density of states for the quantum well structure (solid lines) compared with the density of states for the 'normal' solid (dashed lines).

The halfwidth of this distribution is given by $kT \log_e 2$ or $0.69kT$, which is appreciably smaller than the halfwidth of $1.8kT$ for the three-dimensional case.

A third notable difference between the properties of the bulk semiconductor and the two-dimensional quantum well structure involves the properties of excitons (section 2.4.3). In theory, exciton spectral absorption lines for bulk semiconductors should be visible at an energy just less than that of the bandgap wavelength $\lambda_g (= E_g/hc)$. In practice such absorption lines are only seen in very pure materials, and then usually only at low temperatures. Any impurities present in the semiconductor have the effect of providing screening of the Coulombic interaction between the electrons and holes. In quantum well structures the electrons and holes are much more closely confined so that their wavefunctions overlap more strongly and transition probabilities are higher. In addition, the 'compression' in one dimension has the effect of increasing the exciton binding energy to two or three times that given by eq. (2.28a). Exciton transitions thus become much more pronounced and absorption lines are readily observed even at room temperatures.

If an electric field is now applied along the z direction, that is perpendicular to the plane of the slab, the exciton electron and hole wavefunctions are considerably modified and the exciton resonances both broaden and move to lower energies (ref. 2.15). This is called the quantum confined Stark effect, and is the basis of a type of modulator to be discussed in the next chapter (section 3.9). It should be mentioned that bulk semiconductor materials show a change in absorption at wavelengths close to the bandgap wavelength when an electric field is applied, but the effect is very much smaller than in the quantum well structure. It is also noteworthy that, as materials properties and fabrication techniques improve, quantum wells are increasingly being used in active and passive optical and optoelectronic devices. These include LEDs and lasers (section 5.10.2), photodiodes and waveguides as well as optical modulators. (ref. 2.16)

PROBLEMS

- 2.1 Explain the significance of each of the four quantum numbers used to describe atomic energy states. Show how constraints on the values of these numbers lead to a system of 'shells' and 'subshells' of energy states. Draw up a table of all the possible states for $n = 4$.
- 2.2 Explain what is meant by the Pauli exclusion principle and show, on the basis of your answers to Problem 2.1, what is meant by the electron configurations of the elements. What are the expected electron configurations of diamond ($Z = 6$), Si ($Z = 14$), Ge ($Z = 32$), Sn ($Z = 50$) and Nd ($Z = 60$)? Why might the actual configurations be different from those that you expect?
- 2.3 An electron in an electron microscope is accelerated by a voltage of 25 kV; what is its de Broglie wavelength?
- 2.4 Show that the uncertainty principle can be expressed as $\Delta E \Delta t \geq h/2\pi$, where ΔE and Δt are the uncertainties in energy and time respectively. What is the uncertainty in the velocity of an electron confined in a cube of volume of 10^{-30} m^3 ?

- 2.5 Calculate the energies of the first three levels for a cubic potential well of side length 10^{-10} m. How much energy is emitted if an electron falls from the third to the first level? If the emitted energy is in the form of a photon, what is the frequency of the associated wave?
- 2.6 According to the de Broglie relation we may regard a particle as having a wavelength λ . Show that when an electron is moving along a one-dimensional array of ions with separation a , then a standing wave will be set up when the electron wavevector ($2\pi/\lambda$) is equal to $n\pi/a$ where n is an integer.
- From the standpoint of quantum theory, these standing waves represent electron wavefunctions. In fact two different wavefunctions can be set up corresponding to each n value since each ion may be either a node or an antinode for the standing wave. Show that these two will have different energies. (Hint: consider the Coulombic energy resulting from the interaction of the electronic charge distributions with the ions.) How does this energy difference manifest itself?
- 2.7 At 300 K the conductivity of intrinsic silicon is $5 \times 10^{-4} \Omega^{-1} \text{ m}^{-1}$. If the electron and hole mobilities are 0.14 and $0.05 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ respectively what is the density of electron-hole pairs?
- If the crystal is doped with 10^{22} m^{-3} phosphorus atoms, calculate the new conductivity; repeat for the case of boron doping at the same impurity level. Assume all the impurities are ionized in both cases.
- 2.8 Estimate the ionization energy of donors in GaAs given that $m^* = 0.07m$ and $\epsilon_r = 10.9$.
- 2.9 Calculate the probabilities of finding electrons at energy levels of $E_F + 0.05 \text{ eV}$ and $E_F - 0.05 \text{ eV}$ at $T = 0 \text{ K}$ and 300 K .
- 2.10 The atomic mass number of copper is 63.54 and the density of copper is $8.9 \times 10^3 \text{ kg m}^{-3}$; confirm that its Fermi level is about 7.0 eV.
- 2.11 Calculate the intrinsic carrier concentration in GaAs at 290 K given that the electron effective mass is $0.07m$, the hole effective mass is $0.56m$ and that its energy gap is 1.43 eV.
- 2.12 Calculate the diffusion coefficients of electrons and holes in silicon at 290 K from the data given in Problem 2.7. If it is assumed that the electron and hole lifetimes are both 50 μs , what are their diffusion lengths?
- 2.13 Derive eq. (2.29) for the density of states – use eq. (2.13) to set up a three-dimensional integer space in which each point (i.e. each combination of n_1, n_2, n_3) represents an energy level. Hence show, using eq. (2.11), that the number of energy states with energy less than a reference energy E_R (which equals the appropriate volume of integer space) is $(4\pi/3)(8E_R mL^2/h^2)^{3/2}$. (Remembering that only positive integers n_1, n_2, n_3 have any physical meaning and that there are two spin states per energy level, we arrive at eq. (2.29).) Show also that for two-dimensional integer space, the density of states is independent of the energy.

- 2.14 Using eqs (2.29) and (2.30), derive eq. (2.33) for the electron density in the conduction band of a semiconductor. Make, and if possible justify, the following assumptions: (1) that the energy range of the conduction band can be taken as $E_c \leq E < \infty$; and (2) that we can ignore the term -1 in the denominator of eq. (2.30) as $E - E_F \gg kT$. Furthermore, make the substitution $x = (E - E_c)/kT$ and note that

$$\int_0^{\infty} x^{1/2} \exp(-x) dx = \sqrt{\pi}/2$$

- 2.15 An electron current of 10 mA is injected into a p-type silicon rod of 1 mm² cross-sectional area. Assuming that the excess concentration decreases exponentially and that at 5 mm from the contact the excess concentration has fallen to 40% of its value at the contact, calculate the value at the contact. You may assume that $D_e = 3.4 \times 10^{-3} \text{ m}^2 \text{ s}^{-1}$.
- 2.16 Calculate the equilibrium contact potential for a step junction in silicon if $N_a = 10^{24} \text{ m}^{-3}$ and $N_d = 10^{24} \text{ m}^{-3}$; take $\epsilon_r = 12$, $n_i = 1.5 \times 10^{16} \text{ m}^{-3}$ and $T = 290 \text{ K}$. Also calculate the width of the junction both in equilibrium and with a reverse bias voltage of 50 V.
- 2.17 The resistivities of the p and n materials forming a p-n junction are 4.2×10^{-4} and $2.08 \times 10^{-2} \Omega \text{ m}$ respectively. If the hole and electron mobilities are 0.15 and $0.3 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$, the hole and electron carrier lifetimes are 75 and 150 μs respectively and the intrinsic carrier concentration is $2.5 \times 10^{19} \text{ m}^{-3}$, calculate the saturation current at 290 K given that the junction area is 10^{-6} m^2 . What fraction of the current is carried by holes?
- 2.18 Calculate the maximum wavelengths of light which will give rise to photoeffects in intrinsic GaAs ($E_g = 1.43 \text{ eV}$), CdS ($E_g = 2.4 \text{ eV}$) and InSb ($E_g = 0.225 \text{ eV}$).
- 2.19 Show that the average energy of the free electrons (which is equal to the total energy divided by the electron density) in a solid at 0 K is given by $\langle E \rangle = \frac{3}{5} E_F$. (Hint: the total energy of the free electrons can be calculated by multiplying the population of each energy state by its energy and integrating over all states.)
- 2.20 Show that the most probable electron energy, and therefore the maximum electron density, in the conduction band of a semiconductor is $\frac{1}{2} kT$ above the bottom of the band. Also show that the average electron energy in the conduction band is $\frac{3}{2} kT$, while the halfwidth of the distribution is approximately $1.8 kT$.
- 2.21 Show that an electron of energy E approaching a narrow potential barrier of height V_0 ($V_0 > E$) may penetrate the barrier (i.e. tunnel through it) providing the barrier is not infinitely high. (Hint: if V_0 is not infinite then ψ does not become zero at the barrier; we may use the conditions that ψ and $d\psi/dx$ are continuous at each side of the barrier. Thus ψ must have a value on the far side of the barrier, so that $\psi^* \psi$ (or $|\psi(x)|^2$) is finite, implying that there is some probability of finding the electron beyond the barrier – set up ψ for each of the three regions: before the barrier, within the barrier and beyond the barrier.)

REFERENCES

- 2.1 (a) A. Bar-Lev, *Semiconductors and Electronic Devices*, Prentice Hall, Hemel Hempstead, 1979, Section 6.1.
(b) M. N. Rudden and J. Wilson, *Elements of Solid State Physics* (2nd edn), John Wiley, Chichester, 1993, Chapter 1.
(c) L. Solymar and D. Walsh, *Lectures on the Electrical Properties of Materials* (2nd edn), Oxford University Press, Oxford, 1979, Chapter 4.
(d) B. G. Streetman, *Solid State Electronic Devices* (2nd edn), Prentice Hall, Englewood Cliffs, NJ, 1980.
(e) C. Kittel, *Introduction to Solid State Physics* (5th edn), John Wiley, New York, 1976.
(f) R. A. Smith, *Semiconductors* (2nd edn), Cambridge University Press, Cambridge, 1979.
- 2.2 (a) R. W. Ditchburn, *Light* (2nd edn), Blackie, Glasgow, 1962, Section 4.19 and Appendix IVB.
(b) E. Hecht, *Optics* (2nd edn), Addison-Wesley, Reading, MA, 1987, Chapter 7.
- 2.3 L. Solymar and D. Walsh, *op. cit.*, Chapter 7.
- 2.4 *Ibid.*, pp. 132–6.
- 2.5 A. Bar-Lev, *op. cit.*, Section 6.5.
- 2.6 L. Solymar and D. Walsh, *op. cit.*, pp. 172–3.
- 2.7 B. G. Streetman, *op. cit.*, pp. 58–61.
- 2.8 *Ibid.*, pp. 54–5.
- 2.9 L. Solymar and D. Walsh, *op. cit.*, Section 6.2.
- 2.10 A. Bar-Lev, *op. cit.*, Section 4.2.
- 2.11 (a) M. N. Rudden and J. Wilson, *op. cit.*, Chapter 5.
(b) B. G. Streetman, *op. cit.*, pp. 126–36.
- 2.12 *Ibid.*, pp. 185–93.
- 2.13 (a) J. Goward, *Optical Communication Systems* (2nd edn), Prentice Hall International, Hemel Hempstead, 1995.
(b) D. Wood, *Optoelectronic Semiconductor Devices*, Prentice Hall International, Hemel Hempstead, 1994.
- 2.14 S. M. Sze, *Physics of Semiconductor Devices*, Wiley-Interscience, New York, 1969, Sections 8.3 and 9.3.
- 2.15 E. H. Li and B. L. Weiss, 'Bangap engineering and quantum wells in optoelectronic devices', *Electron. Commun. Eng. J.*, April, 63, 1991.
- 2.16 *Ibid.*

Modulation of light

The advent of the laser (see Chapter 5) and the increasing use of lasers in a wide variety of applications have led to a demand for devices which can modulate a beam of light. Applications of light modulators include wideband analog optical communication systems, switching for digital information recording, information storage and processing, pulse shaping, beam deflection and scanning, and frequency stabilization and *Q*-switching of lasers. Some of these applications are discussed in Chapter 6. In this chapter, we have interpreted the term modulation rather broadly so that we also include sections on scanning and some aspects of laser wavelength tuning. Several of the materials, for example KDP, which are useful in conventional modulators exhibit non-linear effects and consequently may be used for harmonic generation and parametric oscillation. These techniques, together with those described in section 6.5.1.6, enhance the available range of laser wavelengths.

A modulator is a device which changes the irradiance (or direction) of the light passing through it. There are several general types of modulator: namely, mechanical choppers and shutters, passive (or dye) modulators, electro-optic, magneto-optic and elasto-optic (acousto-optic) modulators. The first two types will be covered briefly in section 6.4 in which laser *Q*-switching is discussed. In the remaining types the refractive index and other optical characteristics of a medium are changed by the application of a force field, that is electrical, magnetic or mechanical (acoustical). In these cases, apart from the acousto-optical effect where the variation of refractive index creates a 'diffraction grating', the applied force field modifies the polarizing properties of the medium. This in turn may be used to modify the phase or irradiance of a beam of light propagating through the medium. Accordingly this chapter begins with a brief review of optical polarization, birefringence and optical activity in naturally occurring crystals.

3.1

Elliptical polarization

We have already described (section 1.2.1) plane polarized light in which all of the wave trains comprising a beam of light have their electric vectors lying in the same plane. In many cases of interest a beam of light may consist of two plane polarized wave trains with their planes of polarization at right angles to each other, and which may also be out of phase.

Let us consider initially the special case where the amplitudes of the two wave trains are equal and the phase difference is $\pi/2$. In this case, if the wave trains are propagating in the

z direction, we may write the component electric fields as

$$\mathcal{E}_x = i\mathcal{E}_0 \cos(kz - \omega t)$$

and

$$\mathcal{E}_y = j\mathcal{E}_0 \sin(kz - \omega t)$$

(3.1)

where ω is the angular frequency, $k (= 2\pi/\lambda)$ is the wavenumber or propagation constant and i and j are unit vectors in the x and y directions respectively. The total electric field is the vector sum of the two components, that is

$$\mathcal{E} = \mathcal{E}_x + \mathcal{E}_y$$

or

$$\mathcal{E} = \mathcal{E}_0[i \cos(kz - \omega t) + j \sin(kz - \omega t)] \quad (3.2)$$

The resultant expressed by eq. (3.2) can be interpreted as a single wave in which the electric vector at a given point in space is constant in amplitude but rotates with angular frequency ω . Waves such as this are said to be *circularly polarized*. Figure 3.1 shows the electric field vector (a) at a given instant of time and (b) at a given point in space.

The signs of the terms in eq. (3.2) are such that the electric vector at a given point in space has a clockwise rotation when viewed against the direction of propagation. A wave with such an electric vector is said to be right circularly polarized.

If the sign of the second term is changed (this is equivalent to a change of π in the phase of \mathcal{E}_y) then the sense of rotation is counterclockwise and the wave is said to be left circularly polarized.

When the amplitudes of the electric vectors of the two waves are not the same but the phase difference remains at $\pi/2$ then the resultant electric vector at any point in space rotates at frequency ω but changes in magnitude. The end of the electric vector describes an ellipse as illustrated in Fig. 3.2.

If the component waves can be represented by $\mathcal{E}_x = i\mathcal{E}_0 \cos(kz - \omega t)$ and

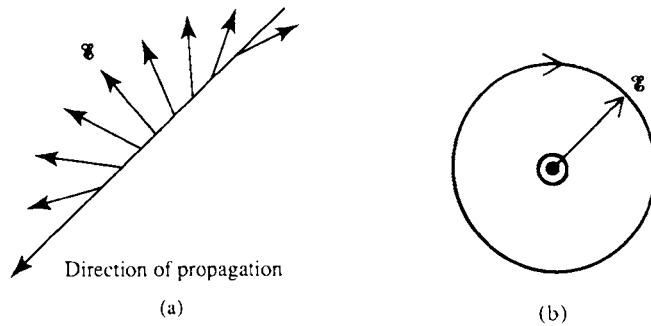


FIG. 3.1 Right circularly polarized light: (a) electric vectors at a given instant in time and (b) rotation of the vector at a given position in space. (Note that in this, and the following diagrams, for the sake of clarity only the electric vectors are shown.)

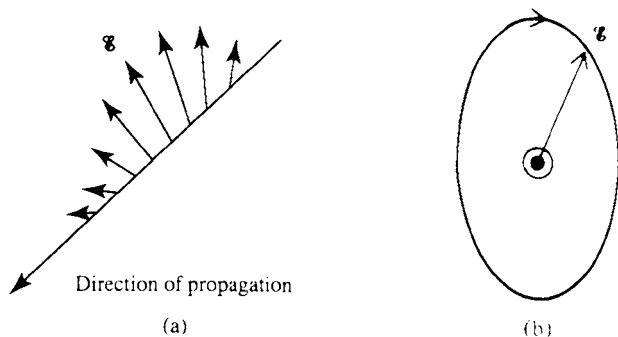


FIG. 3.2 Right elliptically polarized light; electric vectors at (a) a given instant in time and (b) a given position in space.

$\mathcal{E}_y = j\mathcal{E}'_0 \sin(kz - \omega t)$ with $\mathcal{E}_0 \neq \mathcal{E}'_0$, then the major and minor axes of the ellipse are parallel to the x and y axes. In general the electric vector amplitudes are not equal and also there is an arbitrary phase difference ϕ between the component waves. In this case the end of the resultant electric vector again describes an ellipse but with the major and minor axes inclined at an angle of $\frac{1}{2} \tan^{-1} [(2\mathcal{E}_0\mathcal{E}'_0 \cos \phi) / (\mathcal{E}_0'^2 - \mathcal{E}'_0^2)]$ to the x and y axes (see Problem 3.1). In all of these instances, the resultant wave is said to be elliptically polarized and in fact plane and circular polarization are special cases of elliptical polarization. For the purposes of this book it is not instructive to dwell on a full discussion of elliptically polarized light (ref. 3.1) but rather to describe it simply in terms of its components parallel to and perpendicular to a convenient axis or plane.

3.2 Birefringence

Ordinary glass is isotropic in its properties, but many important crystalline optical materials such as calcite (CaCO_3), quartz (SiO_2) and KDP (potassium dihydrogen phosphate, KH_2PO_4) are anisotropic. This anisotropy is due to the arrangement of the atoms being different in different directions through the crystal. Thus, for example, the electric polarization \mathbf{P} produced in the crystal by a given electric field \mathcal{E} is not a simple scalar multiple of the field but in fact varies in a manner that depends on the direction of the applied field in relation to the crystal lattice. One of the consequences of this is that the refractive index, experienced by electromagnetic waves travelling through the crystal, depends on not only the direction of propagation of the waves, but also the direction of the polarization of the waves. Consequently the phase velocity of the waves also depends on these directions. Such crystals are said to be *birefringent* or *doubly refracting*; the names refer to the fact that in general there are *two* different directions of propagation through the crystal that a given incident ray may take depending on the direction of its polarization. The rays corresponding to these directions travel with different velocities and have mutually orthogonal planes of polarization. Consequently, when unpolarized light or light of arbitrary polarization relative to the crystal structure is incident on a doubly refracting crystal, the light propagating through

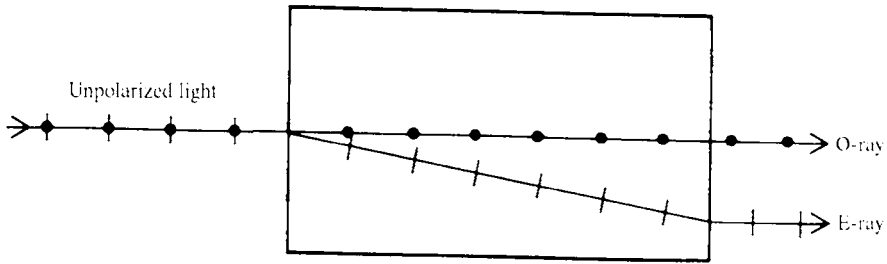


FIG. 3.3 Double refraction by a birefringent crystal.

the crystal may be considered to comprise two independent waves which travel with different velocities. Similarly if light propagates in a given direction through the crystal there are only two possible values of phase velocity, which correspond to two orthogonal states of polarization.

The theory, which is not very straightforward, shows that in general crystals exhibit three different principal refractive indices and two optic axes (see Appendix 2 and ref. 3.2 for further discussion). The optic axes are *directions* in the crystal along which the velocities of the two orthogonally polarized waves are the same. In many important crystals, for example calcite, two of the principal indices are the same and there is only one optic axis. Such crystals are called *uniaxial* whereas other doubly refracting crystals, for example mica, are *biaxial*. In cubic crystals which are isotropic the principal indices are all the same.

A simple way of observing birefringence is to allow a narrow beam of unpolarized light to fall normally onto a parallel-sided calcite plate as shown in Fig. 3.3. The beam is found to divide into two parts. One, the so-called *ordinary* or O-ray (represented in diagrams by •••••), passes straight through the crystal, as might be expected, while the other, the so-called *extraordinary* or E-ray (represented in diagrams by +++++), diverges from the O-ray as it passes through the crystal and then emerges parallel to it. This is found to be the case unless the direction of incidence of the original beam is parallel or perpendicular to the optic axis. The ordinary and extraordinary rays are found to have orthogonal directions of polarization, which are normal to and parallel to the *principal section* of the crystal (shown in Fig. 3.3), which is a plane containing the optic axis and which is normal to a pair of opposite parallel surfaces of the crystal.

The relationship between the different values of refractive index and crystal structure can be shown on the refractive *index ellipsoid* or *optical indicatrix*, which is illustrated in Fig. 3.4 for uniaxial crystals, which displays two refractive indices. As implied above, these are the ordinary refractive index n_o , which does not vary with direction, and the extraordinary refractive index n_e (or more precisely $n_e(\theta)$) whose value depends on the direction of propagation θ relative to the optic axis. Although the index ellipsoid and its applications are described in more detail in Appendix 2, we can illustrate its use by considering light propagating along the direction \mathbf{r} shown in Fig. 3.4, where for convenience (but without loss of generality) the projection of \mathbf{r} on the xy plane is along the y axis. The plane normal to \mathbf{r} which passes through the origin intersects the index ellipsoid along the perimeter of the shaded ellipse shown. The two directions of polarization mentioned above

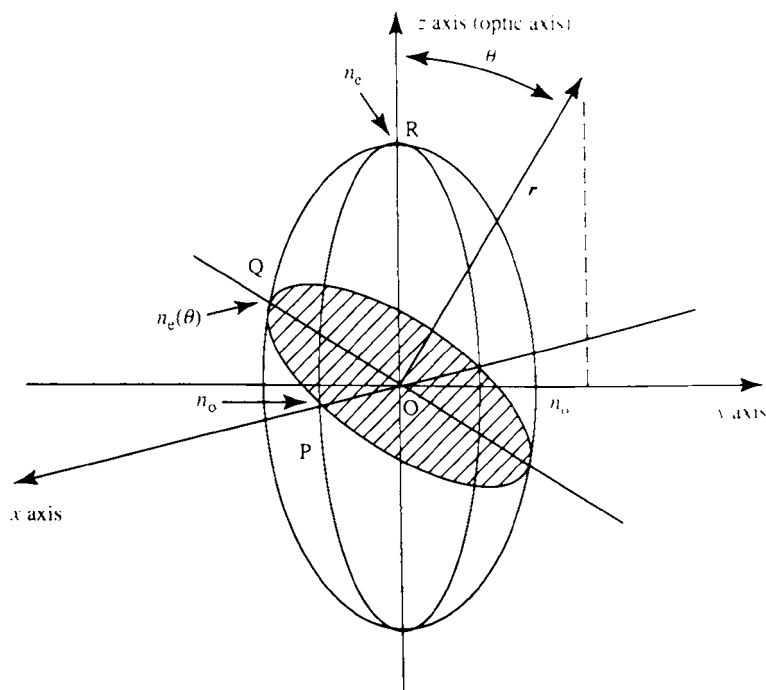


FIG. 3.4 The refractive index ellipsoid for a uniaxial crystal. Waves with polarizations parallel to the x and y axes experience a refractive index n_o , while those polarized parallel to the z axis experience a refractive index n_e . For wave propagation in a general direction r there are two allowed directions of polarization, i.e. parallel to OP with index n_o , and parallel to OQ with index $n_e(\theta)$.

then lie along the principal axes (OP and OQ) of this ellipse. In addition the corresponding refractive indices are given by the magnitudes of OP and OQ . One of these, OP , is independent of θ and represents the refractive index n_o corresponding to the ordinary ray, while the other, OQ , depends on the angle θ and gives the refractive index $n_e(\theta)$ corresponding to the extraordinary ray.

It may help us to visualize the behaviour of doubly refracting materials if we were to draw surfaces in which the distance of each point from the origin corresponds to the values of $n_e(\theta)$ and n_o , for wave propagation in the direction θ . The resulting surfaces would be a sphere for the ordinary rays and an ellipsoid for the extraordinary ones. Such *normal (index) surfaces* for a positive crystal are shown in Fig. 3.5. We can now envisage light spreading out from a point source located at the origin using Huygens' construction. The ordinary rays will form a spherical surface, while the extraordinary ones form an ellipsoidal one, which in the case of positive uniaxial crystals will lie inside the sphere as velocity is inversely proportional to refractive index. The sphere and ellipse will of course coincide as they cross the optic axis, and be separated by the greatest distance, and therefore phase difference, in the direction perpendicular to the optic axis. We may use this construction to help us to envisage the behaviour of phase plates described in the next section.

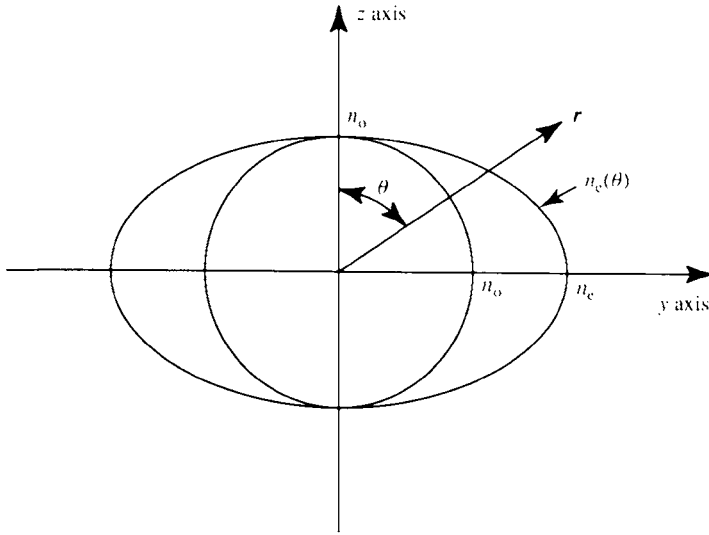


FIG. 3.5 Normal (index) surfaces for a positive uniaxial crystal projected on the yz plane.

3.2.1 Phase plates

Plates of crystals can be cut with particular orientations to the optic axis to produce a desired effect in an optical beam propagating through the plate. For example, plates cut with their large area surfaces parallel to the optic axis can be fabricated to introduce a given phase change between the O- and E- rays as shown in Fig. 3.6. In particular a plate of thickness d such that the optical path difference $|n_o d - n_e d| = \lambda_o/4$ which of course is equivalent to a phase change of $\pi/2$, is called a *quarter-wave plate*. For quartz, for example, d should be equal to 0.0164 mm for sodium light (Problem 3.2). When plane polarized light is incident

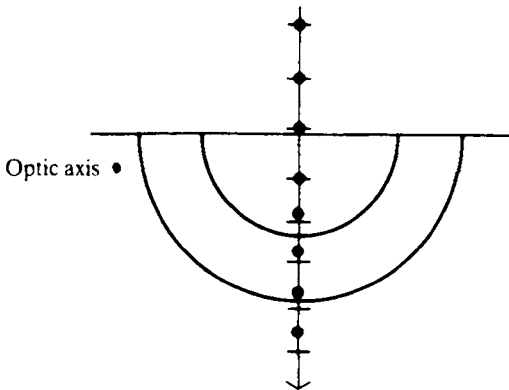


FIG. 3.6 Double refraction by a negative crystal which is cut such that the optic axis is parallel to the crystal surface but normal to the plane of incidence. In this case there is no divergence of the E- and O-rays, but they become increasingly out of phase as they propagate through the crystal.

on a quarter-wave plate, the emergent light is in general elliptically polarized. If the plane of polarization of a plane polarized incident beam, however, is inclined at 45° to the optic axis then the amplitudes of the E and O components will be the same, and the emergent light is circularly polarized as explained in section 3.1. In a similar way, half-wave and whole-wave plates can be fabricated. Such plates are often used in light modulation systems.

3.3 Optical activity

Certain crystals (and liquids) have the ability to rotate the plane of polarization of light passing through them; that is, they are *optically active*. Thus, for example, when a beam of plane polarized light is incident normally on a crystal plate of quartz cut perpendicular to the optic axis, it is found that the emergent beam is also plane polarized but that its electric vector vibrates in a different plane from that of the incident light. The plane of vibration may be rotated in a clockwise sense looking against the oncoming light by *right-handed* or *dextro-rotatory* crystals, or in a counterclockwise sense by *left-handed* or *laevorotatory* crystals. Quartz exists in both forms. It is found that the rotation depends on the thickness of the crystal plate and the wavelength. The rotation produced by a quartz plate 1 mm thick for sodium light is 21.7° while it is 3.67° for 1 mm of sodium chlorate.

Optical activity can be explained by assuming that in optically active crystals the velocity of propagation of circularly polarized light is different for different directions of rotation, that is the crystal has refractive indices n_r and n_l for right and left circularly polarized light. It is easy to show (see Problem 3.3) that a plane polarized wave can be resolved into two circularly polarized waves with opposite directions of rotation. If these travel through the crystal at different speeds, a phase difference will be introduced between them at different distances through the crystal. This corresponds to a rotation of the plane of the plane polarized wave which results from the recombination of the two circularly polarized waves.

3.4 Electro-optic effect

When an electric field is applied across an optical medium the distribution of electrons within it is distorted so that the polarizability and hence the refractive index of the medium changes anisotropically. The result of this electro-optic effect may be to introduce new optic axes into naturally doubly refracting crystals, for example KDP, or to make naturally isotropic crystals, for example gallium arsenide, doubly refracting.

The change in refractive index as a function of the applied field can be obtained from an equation of the form (see ref. 3.3 and Appendix 2)

$$\Delta(1/n^2) = r\mathcal{E} + P\mathcal{E}^2 \quad (3.3)$$

where r is the *linear* electro-optic coefficient and P is the *quadratic* electro-optic coefficient. In solids, the linear variation in the refractive index associated with $r\mathcal{E}$ is known as the Pockels effect while the variation arising from the quadratic term is called the Kerr effect (not to be confused with the magneto-optic effect also named after Kerr).

In the case of the Pockels effect, the precise effects of the applied electric field depend on the crystal structure and symmetry of the material under consideration. With KDP, for example, if the electric field is applied along the z direction then the x and y principal axes are rotated through 45° into new principal axes x' and y' and the refractive indices in these new directions become (see Appendix 2)

$$\begin{aligned} n_{x'} &= n_o + \frac{n_o^3}{2} r_{63} \mathcal{E}_z \\ n_{y'} &= n_o - \frac{n_o^3}{2} r_{63} \mathcal{E}_z \end{aligned} \quad (3.4)$$

Thus,

$$\Delta\left(\frac{1}{n^2}\right) = -\frac{2\Delta n}{n^3} = r_{63} \mathcal{E}_z \quad (3.5)$$

which is in agreement with eq. (3.3) assuming that the Kerr constant P is very small, where r_{63} is the appropriate electro-optic coefficient for KDP. In the interests of notational convenience we shall drop the subscripts from r , though it should be remembered that the precise coefficient used depends on the crystal symmetry and directions of the field and light wave.

EXAMPLE 3.1 Change in refractive index due to the Pockels effect

We may calculate the change in refractive index for a 10 mm wide crystal of deuterium-substituted KDP (i.e. KD*P) for an applied voltage of 4000 V.

From eq. (3.5) and using the data in Table 3.1 below we have

$$n - n_o = \pm \frac{1}{2} \times 26.4 \times 10^{-12} \times 1.51^3 \times 4000/10^{-2}$$

or

$$|n - n_o| = 1.8 \times 10^{-5}$$

Let us consider a beam of plane polarized light propagating in the z direction through a crystal such as KDP with its plane of polarization at 45° to the induced axes x' and y' as shown in Fig. 3.7. If the incident beam is given by $\mathcal{E} = \mathcal{E}_0 \cos(\omega t - kz)$, then the components along the x' and y' directions will be

$$\begin{aligned} \mathcal{E}_{x'} &= \frac{\mathcal{E}_0}{\sqrt{2}} \cos(\omega t - kz) \\ \mathcal{E}_{y'} &= \frac{\mathcal{E}_0}{\sqrt{2}} \cos(\omega t - kz) \end{aligned} \quad (3.6)$$

In view of the fact that these components experience the refractive indices given by eq. (3.4), they will become increasingly out of phase as they propagate through the crystal. Thus if the crystal is of thickness L the phase changes of the two components will be

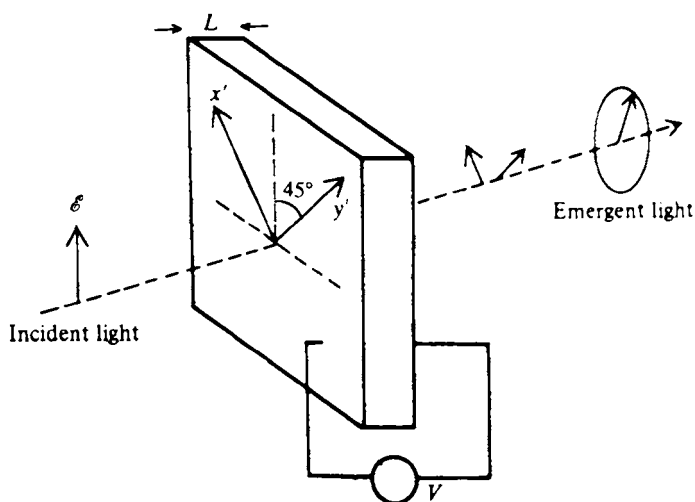


FIG. 3.7 A beam of plane polarized light incident on an electro-optic crystal plate subject to a voltage V will be resolved into components with polarizations along x' and y' , the induced principal directions. The induced birefringence occurs in a plane normal to the applied electric field.

$$\phi_{x'} = \frac{2\pi}{\lambda_0} n_x L$$

$$\phi_{y'} = \frac{2\pi}{\lambda_0} n_y L$$

Using eqs (3.4) which relate the refractive index to the electric field, we see that

$$\begin{aligned} \phi_{x'} &= \frac{2\pi}{\lambda_0} L n_o \left(1 + \frac{1}{2} r n_o^2 \mathcal{E}_z\right) \\ \phi_{y'} &= \frac{2\pi}{\lambda_0} L n_o \left(1 - \frac{1}{2} r n_o^2 \mathcal{E}_z\right) \end{aligned} \quad (3.7)$$

which may be written

$$\begin{aligned} \phi_{x'} &= \phi_0 + \Delta\phi \\ \phi_{y'} &= \phi_0 - \Delta\phi \end{aligned} \quad (3.7a)$$

where

$$\Delta\phi = \frac{\pi}{\lambda_0} L r n_o^3 \mathcal{E}_z = \frac{\pi}{\lambda_0} r n_o^3 V \quad (3.8)$$

In eq. (3.8) we have taken \mathcal{E}_z to equal V/L , where V is the applied voltage.

The net phase shift, or total *retardation*, between the two waves resulting from the

application of the voltage V is seen to be

$$\Phi = \phi_{x'} - \phi_{y'} = 2\Delta\phi = \frac{2\pi}{\lambda_0} r n_o^3 V \quad (3.9)$$

and the emergent light will in general be elliptically polarized.

From eqs (3.6) and (3.7), the components of the wave emerging from the electro-optic crystal can now be written as (omitting common phase factors)

$$E_{x'} = \frac{E_0}{\sqrt{2}} \cos(\omega t + \Delta\phi) \quad (3.10)$$

and

$$E_{y'} = \frac{E_0}{\sqrt{2}} \cos(\omega t - \Delta\phi) \quad (3.10a)$$

The phase shift $\Delta\phi$ for each component depends directly on the applied voltage V (eq. 3.8) so that we can vary the phase shift by varying the voltage applied to a given crystal. Suppose that we now insert a plane polarizing element orientated at right angles to the polarizing element producing the original plane polarized beam after the electro-optic crystal, as shown in Fig. 3.8. Then, as we can see from Fig. 3.8, the transmitted electric field components will be $-E_{x'}/\sqrt{2}$ and $E_{y'}/\sqrt{2}$. That is, using eqs (3.10) we can write the

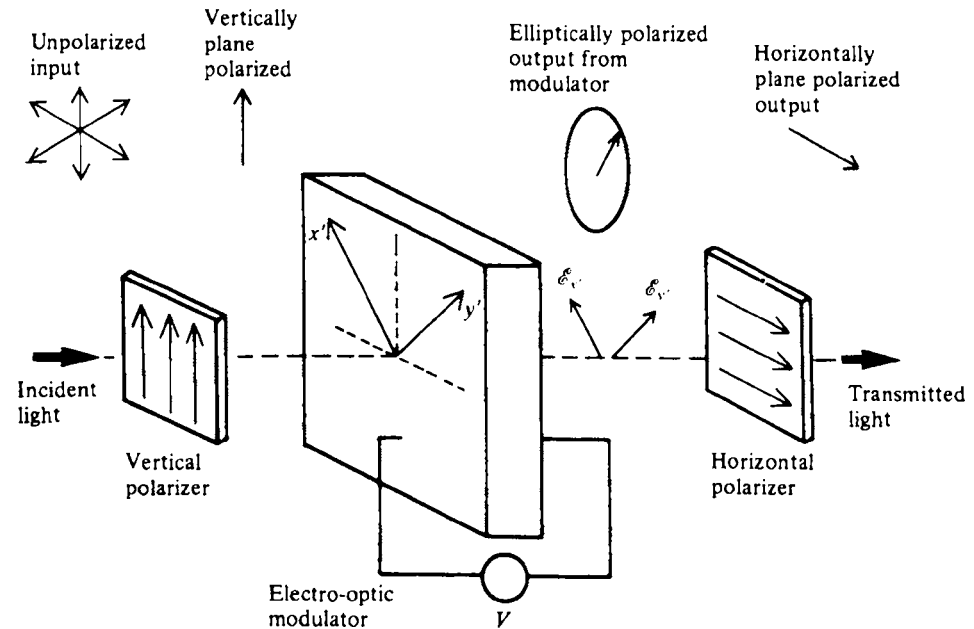


FIG. 3.8 Arrangement of the components of a Pockels electro-optic modulator in which an electro-optic crystal is placed between crossed polarizers. The state of polarization at various positions within the device is also indicated; the components transmitted by the horizontal polarizer are $-E_{x'}/\sqrt{2}$ and $E_{y'}/\sqrt{2}$.

transmitted electric field as

$$\mathcal{E} = -\frac{\mathcal{E}_0}{2} [\cos(\omega t + \Delta\phi) - \cos(\omega t - \Delta\phi)]$$

or

$$\mathcal{E} = -\mathcal{E}_0 \sin \Delta\phi \sin \omega t$$

Thus the irradiance of the transmitted beam, which is given by averaging \mathcal{E}^2 over a complete period $T = 2\pi/\omega$, can be written as

$$I = \frac{\omega}{2\pi} \int_0^{2\pi/\omega} \mathcal{E}^2 dt$$

or

$$I = I_0 \sin^2 \Delta\phi = I_0 \sin^2(\Phi/2) \tag{3.11}$$

where I_0 is the irradiance of the light incident on the electro-optic crystal. As the phase retardation in the Pockels effect is proportional to the voltage, we can see from eqs (3.11) and (3.9) that the transmittance as a function of applied voltage is given by

$$\frac{I}{I_0} = \sin^2\left(\frac{\pi}{\lambda_0} r n_o^3 V\right)$$

which we may write as

$$\frac{I}{I_0} = \sin^2\left(\frac{\pi}{2} \frac{V}{V_\pi}\right) \tag{3.12}$$

where $V_\pi (= \lambda_0/(2rn_o^3))$ is the voltage required for maximum transmission, that is $I = I_0$. V_π is often called the *half-wave voltage* since it causes the two waves polarized parallel to the

TABLE 3.1 Characteristics of some electro-optic materials used in Pockels cells

Material	Linear electro-optic coefficient, r (pm V ⁻¹)	n_o ,†	n_e , †	Relative permittivity,‡ ϵ_r
KH ₂ PO ₄ (KDP)	10.6	1.51	1.47	42
KD ₂ PO ₄ (KD*P)	26.4	1.51	1.47	50
AH ₂ PO ₄ (ADP)	8.5	1.52	1.48	12
Cadmium telluride (CdTe)	6.8	2.6		7.3
Lithium tantalate (LiTaO ₃)	30.3	2.175	2.180	43
Lithium niobate (LiNbO ₃)	30.8	2.29	2.20	18
Gallium arsenide (GaAs)	1.6	3.6		11.5
Zinc sulfide (ZnS)	2.1	2.32		16
Barium borate (BaB ₂ O ₄)		1.67	1.56	7.4

† Values near a wavelength of 550 nm.

‡ Low frequency values.

Note: Several of the materials listed have more than one linear electro-optic coefficient; we have quoted the one which is relevant for use in Pockels cell modulators. (See ref. 3.3b, p. 112, for further details.)

principal axes to acquire a relative spatial displacement of $\lambda_0/2$, which is equivalent to a phase difference of π . Thus a beam of plane polarized light incident on the modulator would have its plane of polarization rotated by 90° when a voltage V_π is applied to the modulator (see Example 3.2). The value of V_π depends on the electro-optic material and the wavelength (see Problem 3.4 and Table 3.1).

Thus we see that the transmittance of the system shown in Fig. 3.8 can be altered by the application of a voltage along the direction of propagation as illustrated in Fig. 3.9. Such systems are referred to as Pockels electro-optic modulators.

It is obvious that the modulation is not linear; indeed from eq. (3.12) for small voltages V , the transmitted irradiance is proportional to V^2 . The effectiveness and ease of operation of a Pockels modulator can be enhanced by including a quarter-wave plate in the beam between the initial polarizer and the modulator as shown in Fig. 3.10(a). This introduces a phase difference of $\pi/2$ between the two polarized components before they enter the voltage-sensitive modulator. A bias is therefore introduced into the transmission curve so that the transmission is varied about the point Q, as illustrated in Fig. 3.10(b), rather than about zero. The change in transmission in the vicinity of Q is more nearly linear with voltage than at the origin.

EXAMPLE 3.2 Half-wave voltage

We may calculate the half-wave voltage for KDP, for example, at a wavelength of $1.06 \mu\text{m}$, using Table 3.1. We have

$$V_\pi = \frac{\lambda_0}{2rn_o^3} = \frac{1.06 \times 10^{-6}}{2 \times 10.6 \times 10^{-12} \times (1.51)^3}$$

$$V_\pi = 14.5 \text{ kV}$$

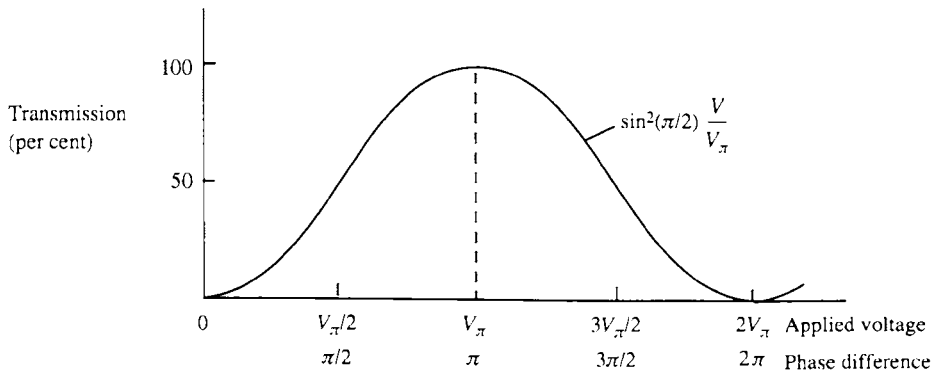


FIG. 3.9 Transmission curve for the system shown in Fig. 3.8 as a function of the applied voltage.

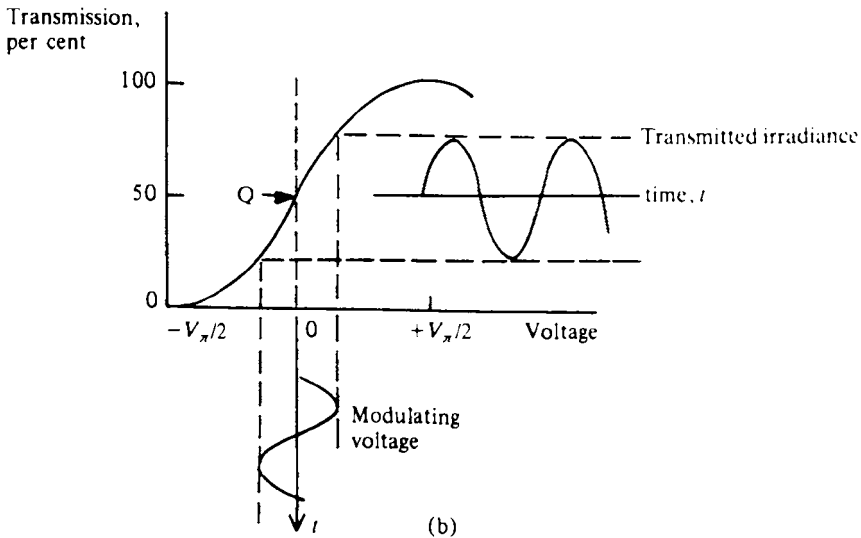
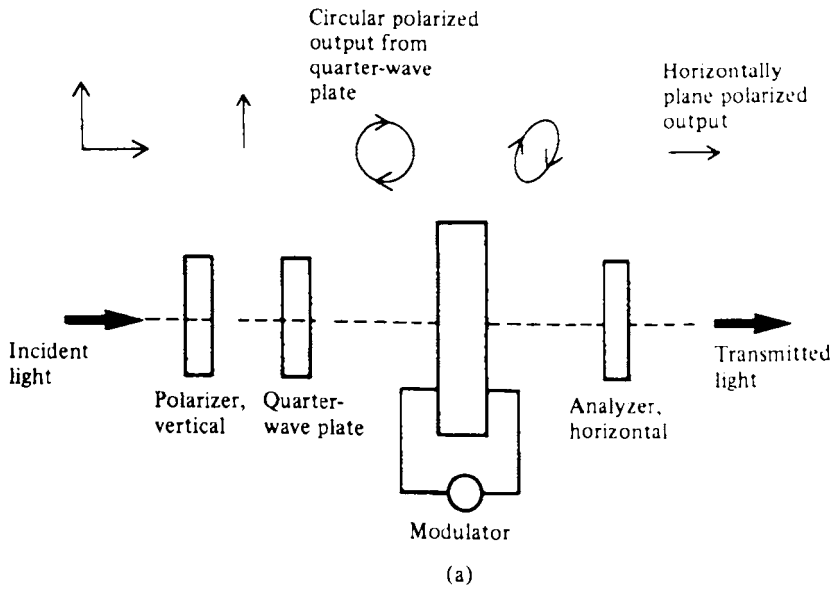


FIG. 3.10 Arrangement of the components of a Pockels electro-optic cell biased with a quarter-wave plate (a) and the resulting transmission as a function of applied voltage (b). The bias results in a 50% irradiance transmission; in the vicinity of this point the variation of transmission with applied voltage is almost linear.

With a quarter-wave plate bias, we see that the phase difference between the components is

$$\Phi = \frac{\pi}{2} + 2\Delta\phi = \frac{\pi}{2} + \pi \frac{V}{V_\pi}$$

and therefore from eq. (3.12)

$$\frac{I}{I_0} = \sin^2\left(\frac{\pi}{4} + \frac{\pi}{2} \frac{V}{V_\pi}\right) = \frac{1}{2} \left(1 + \sin \frac{\pi V}{V_\pi}\right)$$

For small values of applied voltage V (up to about 5% of V_π), $\sin(\pi V/V_\pi) = \pi V/V_\pi$ and the change in irradiance is therefore nearly linear with V . If, therefore, a small sinusoidally varying voltage of amplitude V_0 and frequency f is applied to the modulator, then the irradiance of the transmitted beam will also vary at frequency f , as illustrated in Fig. 3.10(b). That is, we may write

$$\frac{I}{I_0} = 0.5 + \frac{\pi V_0}{2} \sin 2\pi f t \quad (3.13)$$

where $V_0 \sin 2\pi f t = V/V_\pi$ should be very much less than unity otherwise the irradiance variation will be distorted and contain an appreciable amount of higher order harmonics.

The modulator described above is called a longitudinal effect device as the electric field is applied in the direction of propagation of the beam. This can be done by using electrodes with small apertures in them on either side of the electro-optic crystal (Fig 3.11), or by evaporating semitransparent conducting films onto the crystal surfaces. Both of these techniques suffer from obvious disadvantages. To avoid these, an electro-optic modulator with a cylindrical crystal and ring electrode geometry has been developed. This device, which is shown in Fig. 3.11(b), results in very uniform transmission (or polarization) across the effective aperture of the device.

Alternatively, we can use the transverse mode of operation in which the field is applied normal to the direction of propagation. In this case the field electrodes do not interfere with the beam and the retardation (or phase difference), which is proportional to the electric field multiplied by the crystal length, can be increased by the use of longer crystals. (In the longi-

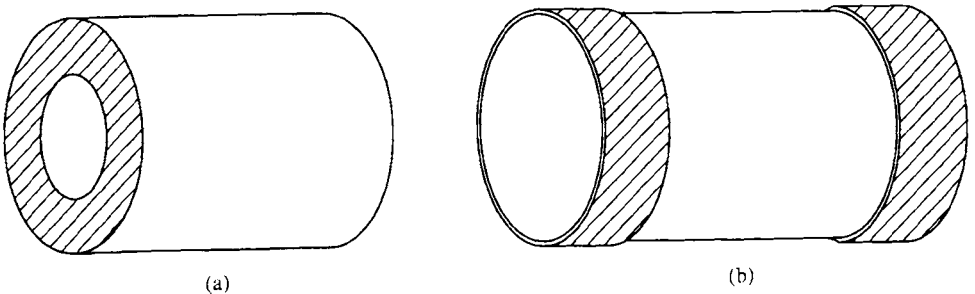


FIG. 3.11 Longitudinal electro-optic cells: (a) shows a cell with end electrodes of relatively small aperture; (b) shows the ring electrode geometry with larger aperture. Typical cell dimensions are length 25 mm, radius 6 mm and electrode width 8 mm.

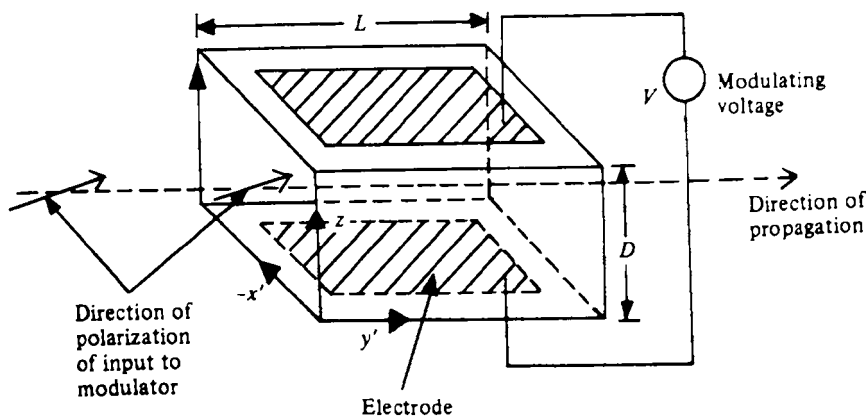


FIG. 3.12 Transverse electro-optic modulator. The electric field is applied normal to the direction of propagation.

tudinal effect the retardation is independent of crystal length.) Suppose that as before the applied field is in the z direction while the direction of propagation is along the y' -induced principal axis, as shown in Fig. 3.12. Then, if the incident light is polarized in the $x'z$ plane at 45° to the x' principal axis, the retardation, using eq. (A2.15) and $n_z = n_e$ from Appendix 2, is

$$\begin{aligned}
 \Delta\phi &= \phi_{x'} - \phi_z = \frac{2\pi}{\lambda_0} L(n_{x'} - n_z) \\
 &= \frac{2\pi}{\lambda_0} L \left((n_o - n_e) + \frac{n_o^3}{2} r' \mathcal{E}_z \right) \\
 &= \frac{2\pi L}{\lambda_0} (n_o - n_e) + \frac{\pi}{\lambda_0} n_o^3 \frac{VL}{D}
 \end{aligned} \tag{3.14}$$

where L is the length of the crystal, D is the crystal dimension in the direction of the applied voltage V and n_o, n_e are the refractive indices for light polarized parallel to the principal directions. The voltage-independent term will bias the irradiance transmission curve. The half-wave voltage may be reduced by having a long, thin cell. Therefore the frequency response of transverse cells is better than in longitudinal cells as it is easier to change small voltages. However, transverse modulators suffer from having very small apertures.

In many practical situations the modulation signal is at very high frequencies and may occupy a large bandwidth so that the wide frequency spectrum available with laser sources may be fully utilized. The capacitance of the modulator and finite transit time of the light through it give rise to limitations in the bandwidth and maximum modulation frequency. Let C be the capacitance due to the electro-optic crystal and its electrodes and R_s be the internal resistance of the modulating source. Then if R_s is greater than $(2\pi f_0 C)^{-1}$, where f_0 is the average modulation frequency, most of the modulation potential drop will be across R_s and therefore will be wasted as it will not contribute to the electro-optic retardation. This problem can be overcome by connecting the crystal in a resonant circuit as shown in Fig. 3.13. The

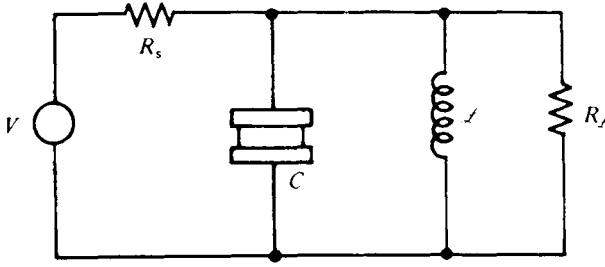


FIG. 3.13 Electro-optic crystal represented by a parallel plate capacitor C connected in a resonant circuit.

value of the inductance \mathcal{L} is such that $4\pi^2 f_0^2 = 1/\mathcal{L}C$ so that at resonance ($f=f_0$) the impedance of the circuit is simply R_j , which is chosen to be greater than R_s and hence most of the modulation voltage appears across the crystal. The resonant circuit has a finite bandwidth, that is its impedance is high only over the frequency range $\Delta f = (2\pi R_j C)^{-1}$ (centred on f_0). Therefore the maximum modulation bandwidth must be less than Δf for the modulated signal to be a faithful representation of the applied modulating voltage.

In practice the bandwidth Δf is governed by the specific application, though bandwidths in the region of 10^8 – 10^9 Hz are readily obtained. In addition, if a peak phase difference or retardation is required we can evaluate the power we need to apply to the crystal. The peak retardation $\Phi_m = (2\pi/\lambda_0) r n_o^3 V_m$ (eq. 3.9) corresponds to a peak modulating voltage $V_m = (\epsilon_z)_m L$. The power $P = V_m^2/2R_j$ needed to obtain the peak retardation is therefore related to the modulation bandwidth by

$$P = \frac{\Phi_m^2 \lambda_0^2 C 2\pi \Delta f}{2(4\pi^2 r^2 n_o^6)}$$

or

$$P = \frac{\Phi_m^2 \lambda_0^2 A \epsilon_r \epsilon_0 \Delta f}{4\pi r^2 n_o^6 L} \quad (3.15)$$

We have taken the capacitance of the crystal at the modulation frequency f_0 to be $C = A\epsilon_r\epsilon_0/L$, where A is the cross-sectional area of the crystal normal to the direction of propagation, in which the crystal length is L .

EXAMPLE 3.3 Power requirement for modulation using a Pockels cell

We may estimate the power required to give a phase retardation of $\pi/30$ at a frequency bandwidth of 10^9 Hz using a KD*P Pockels cell with a circular aperture of 25 mm diameter and 30 mm length at a wavelength of 633 nm.

From eq. (3.15) and using Table 3.1 we find that

$$P = \frac{\pi^2 \times (633 \times 10^{-9})^2 \times \pi \times (12.5 \times 10^{-3})^2 \times 50 \times 8.85 \times 10^{-12} \times 10^9}{30^2 \times 4 \times \pi \times (26.4 \times 10^{-12})^2 \times (1.51)^6 \times 30 \times 10^{-3}} \\ = 306.4 \text{ W}$$

The maximum modulation frequency f_m should be such that the electric field applied to the crystal does not change substantially in a time equal to the transit time t_t of the light through the crystal; that is,

$$t_t = Ln/c \ll 1/f_m$$

Typically $L = 10$ mm and hence f_m must be substantially less than 2×10^{10} Hz in KDP where the refractive index $n \approx 1.5$.

To overcome this restriction the modulating signal can be applied transversely in the form of a wave travelling along the electrodes with a velocity equal to the phase velocity of the optical signal propagating through the modulating crystal. The optical wave then experiences a constant refractive index as it passes through the modulator and much higher modulation frequencies are possible. Although in principle it is quite easy to arrange synchronization of the electrical and optical waves, in practice it is more difficult because of the limitations of the optical materials available. Ideally we need a material for which $n = \sqrt{\epsilon_r}$, where ϵ_r is the relative permittivity (or dielectric constant) of the medium. In most materials this is not the case ($n < \sqrt{\epsilon_r}$) and the desired synchronization must be obtained by reducing ϵ_r by including air gaps in the electrical waveguide cross-section. Alternatively, a travelling wave modulator may be realized by, in effect, slowing down the optical wave. This may be achieved, for example, by letting it propagate through the modulator along the zigzag path shown in Fig. 3.14.

3.4.1 Materials

Any transparent crystal lacking a centre of symmetry exhibits a first-order electro-optic effect. To be useful, however, such crystals must have a sizeable electro-optic coefficient r and be available in reasonably sized, good quality crystals at modest cost. Some of the properties of technologically useful materials are listed in Table 3.1 above.

Potassium dihydrogen phosphate and ammonium dihydrogen phosphate crystals (KDP

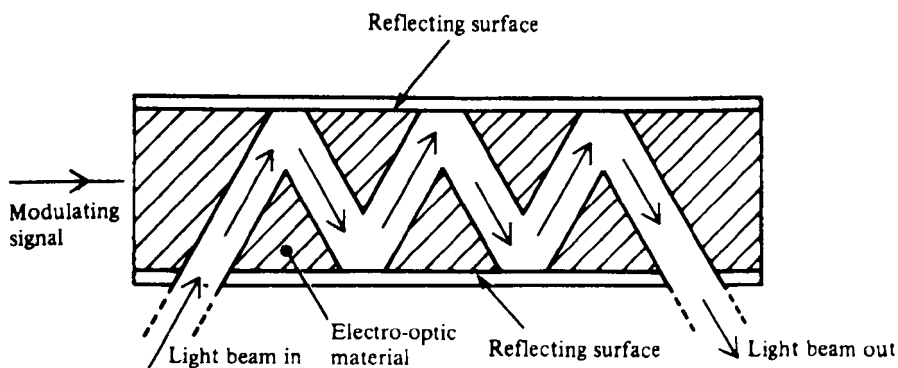


FIG. 3.14 Travelling wave modulator. The light beam follows the zigzag path shown so that in effect it propagates along the modulator at the same speed as the modulating signal and thereby experiences a constant modulating field.

and ADP) are available in large sizes at relatively low cost but are very hygroscopic and fragile. They also have rather large half-wave voltages V_{π} ; however, if deuterium is substituted for hydrogen (i.e. KD*P) the electro-optic properties are greatly enhanced. Other materials such as lithium tantalate and lithium niobate (LiTaO_3 and LiNbO_3) have much smaller half-wave voltages, but reasonably sized crystals are rather expensive. Cadmium telluride and gallium arsenide are used in the infrared in the spectral ranges 1–28 μm and 1–14 μm respectively.

3.5 Kerr modulators

Many isotropic media, both solids and liquids, when placed in an electric field behave as uniaxial crystals with the optic axis parallel to the electric field. In this optoelectric effect, which was discovered in glass by J. Kerr in 1875, the change in refractive index is proportional to the square of the applied field as we indicated in section 3.4. The difference in refractive indices for light polarized parallel to and perpendicular to the induced optic axis is given by

$$\Delta n = n_p - n_s = K\lambda_0 \mathcal{E}_2 \quad (3.16)$$

where K is the Kerr constant (see Table 3.2).

The electric field induces an electric moment in non-polar molecules and changes the moment of polar molecules. There is then a reorientation of the molecules by the field which causes the medium as a whole to become anisotropic. This explains the delay that occurs between the application of the field and the appearance of the maximum effect. The delay can be several seconds, but for non-polar liquids the delay is very small and probably less than 10^{-11} s. A Kerr cell filled with one of these liquids and placed between crossed polaroids can be used as an optical switch or modulator instead of a Pockels cell. Modulation at frequencies up to 10^{10} Hz has been obtained.

Although liquid Kerr cells containing nitrobenzene have been used extensively for many years, for example in the accurate measurement of the velocity of light, they suffer from the disadvantage of requiring a large power to operate them. A more promising approach is to use mixed ferroelectric crystals operating at a temperature near to the Curie point where a greatly enhanced optoelectric effect is observed. (Ferroelectric materials exhibit a spontaneous electric polarization, similar to the spontaneous magnetization of ferromagnetic materials, below a certain temperature. This is the Curie point, above which the crystal structure

TABLE 3.2 Typical values of the Kerr constant
 K for $\lambda = 589.3$ nm at about 20°C

Material	$K (\times 10^{-14} \text{ m V}^{-2})$
Water	5.2
Nitrobenzene	244
Nitrotoluene	137
Classes – various	0.03–0.17

changes and the ferroelectricity disappears.) Potassium tantalate niobate (KTN), which is used in Kerr effect devices, is a mixture of two crystals with high and low Curie points giving a Curie point for the mixture near to room temperature. The crystal has to be 'poled' by the application of a large bias voltage. This has the effect of causing the ferroelectric domains with electric polarization in the direction of the applied field to grow at the expense of the other domains until the whole crystal is polarized in one direction. This then reduces the a.c. voltage required for 100% modulation to about 50 V peak and the half-wave voltage is very much less than for other materials, being about 250 V in KTN and barium titanate (BaTiO_3). Nevertheless, most practical electro-optic modulators make use of the Pockels effect.

3.5.1 Optical frequency Kerr effect

In the optical frequency Kerr effect, as the name implies, the refractive index change is brought about by an applied optical frequency field. This effect offers the interesting possibility of one beam of light being used to switch another if the refractive index change can be used to eliminate the second beam. Such optical beam switching can be achieved using a Fabry-Perot interferometer with an electro-optic material as the spacer between the reflecting plates. Intense maxima in the interference pattern formed by light passing through the interferometer occur when $p\pi = (2\pi/\lambda_0)nd \cos \theta$, where p is the order of interference, d is the plate separation and θ the angle between an internally reflected ray and the surface normal. A small change in the refractive index n of the spacer material induced by an optical beam would detune the interferometer and switch off the transmitted beam.

3.6

Scanning and switching

We saw in section 3.4 that the application of a voltage V_π to a Pockels cell will in effect rotate the plane of polarization of the transmitted optical beam through 90° . Thus if a block of birefringent material is placed after the cell the beam can be switched from one position to another as shown in Fig. 3.15. An array of m such combinations in sequence can obviously be used to address 2^m different locations. Such a system may be used, for example, in bit-oriented optical memories (ref. 6.4d, Chapter 21).

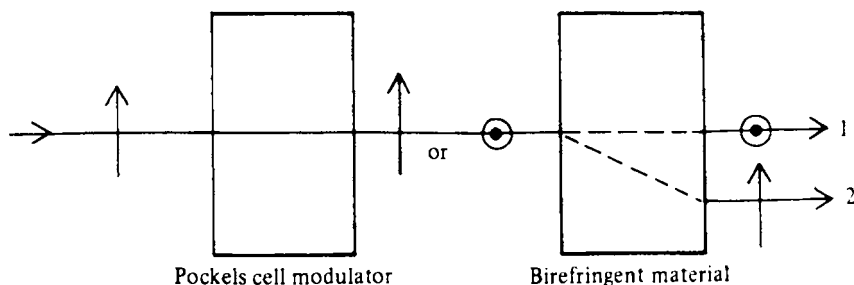


FIG. 3.15 Beam switching using a Pockels cell modulator. As the applied voltage is changed from zero to V_π the beam is switched from position 1 to position 2.

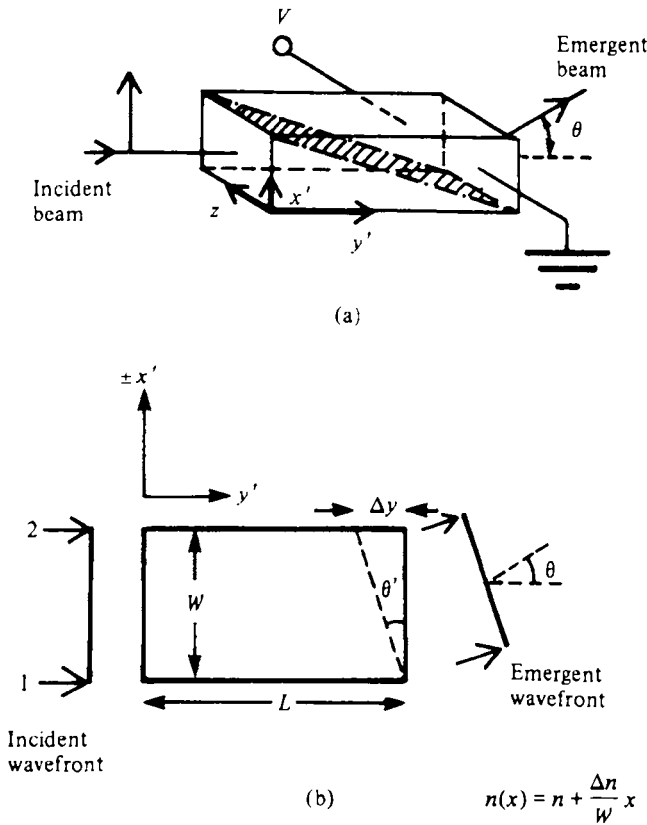


FIG. 3.16 Schematic diagram of a beam deflector: (a) the double prism KDP beam deflector and (b) the principle of beam deflection in a medium where the refractive index varies linearly in a direction normal to the direction of propagation.

Alternatively the arrangement shown in Fig. 3.16(a) can be used for beam switching. Here we have two similar prisms of KDP, for example, but with opposite orientations. Thus, if an electric field is applied in the z direction and the optical beam travels in the direction of one of the induced principal axes with its polarization parallel to the other principal axis, then the beam will 'see' different refractive indices in the two prisms. The difference in refractive indices will be $n_0^3 r' \mathcal{E}_z$; that is, from eqs (3.4), we see that a ray entirely in the upper prism travels in a medium of refractive index

$$n_2 = n_0 - \frac{n_0^3}{2} r' \mathcal{E}_z$$

while a ray entirely in the lower prism travels in a medium where the effect of the applied field is reversed so that

$$n_1 = n_0 + \frac{n_0^3}{2} r' \mathcal{E}_z$$

We can see that this arrangement leads to a deflection of the beam by considering Fig. 3.16(b). Here we show a crystal whose refractive index and hence optical path length varies with the transverse distance x across the crystal. If we assume that the variation of n with x is uniform then ray 1 'sees' a refractive index n while ray 2 'sees' an index $n + \Delta n$. The rays 1 and 2 will traverse the crystal in times t_1 and t_2 where

$$t_1 = \frac{Ln}{c} \quad \text{and} \quad t_2 = \frac{L(n + \Delta n)}{c}$$

The difference in transit times results in ray 2 lagging behind ray 1 by a distance $\Delta y = L\Delta n/n$; this is equivalent to a deflection of the wavefront by an angle θ' measured inside the crystal just before the beam emerges. We can see from Fig. 3.16(b) that $\theta' = \Delta y/W$, where W is the width of the crystal.

Using Snell's law, the angle of beam deflection θ measured outside the crystal is given by (assuming θ is small)

$$\theta = n\theta' = n\Delta y/W$$

or

$$\theta = L\Delta n/W$$

Thus, using the arrangement shown in Fig. 3.16(a), the deflection θ is given by

$$\theta = (L/W)n_o^3 r' \mathcal{E}_z \quad (3.17)$$

Light beams can also be deflected by means of diffraction gratings which are electro-optically induced in a crystal by evaporating a periodic metallic grating electrode onto the crystal. A voltage applied to this electrode induces a periodic variation in the refractive index thereby creating an efficient phase diffraction grating. This technique is especially useful in the integrated optical devices discussed in Chapter 9. Beam deflectors and scanners are used in laser displays, printers and scribes, for optical data storage systems and in optical character recognition.

3.7 Magneto-optic devices

The presence of magnetic fields may also affect the optical properties of some substances thereby giving rise to a number of useful devices. In general, however, as electric fields are easier to generate than magnetic fields, electro-optic devices are usually preferred to magneto-optic devices.

3.7.1 Faraday effect

This is the simplest magneto-optic effect and the only one of real interest for optical modulators; it concerns the change in refractive index of a material subjected to a steady magnetic field. Faraday found in 1845 that when a beam of plane polarized light passes through a substance subjected to a magnetic field, its plane of polarization is observed to rotate by

TABLE 3.3 Typical values of the Verdet constant V for $\lambda = 589.3$ nm

Material	V (rad m ⁻¹ T ⁻¹)
Quartz (SiO ₂)	4.0
Zinc sulfide (ZnS)	82
Crown glass	6.4
Flint glass	23
Sodium chloride (NaCl)	9.6

an amount proportional to the magnetic field component parallel to the direction of propagation. This is very similar to optical activity which, as we saw in section 3.3, results from certain materials having different refractive indices n_r and n_l for right and left circularly polarized light. There is one important difference in the two effects. In the Faraday effect the sense of rotation of the plane of polarization is independent of the direction of propagation. This is in contrast to optical activity where the sense of rotation is related to the direction of propagation. Thus in the case under discussion the rotation can be doubled by reflecting the light back through the Faraday effect device.

The rotation of the plane of polarization is given by

$$\theta = VBL \quad (3.18)$$

where V is the Verdet constant (see Table 3.3 for some representative values), B is the magnetic flux density parallel to the direction of propagation and L is the path length in the material. The Faraday effect is small and wavelength dependent; the rotation for dense flint glass is $\theta \approx 1.6^\circ \text{ mm}^{-1} \text{ T}^{-1}$ at $\lambda_0 = 589.3$ nm.

We can also express θ in terms of the refractive indices n_r and n_l , that is

$$\theta = \frac{\pi}{\lambda_0} (n_r - n_l)L$$

A Faraday rotator used in conjunction with a pair of polarizers acts as an optical isolator which allows a light beam to travel through it in one direction but not in the opposite one. It may therefore be used in laser amplifying chains to eliminate reflected, backward-travelling waves, which are potentially damaging. The construction of a typical isolator is shown in Fig. 3.17.

Light passing from left to right is polarized in the vertical plane by polarizer P_1 . The Faraday rotator is adjusted to produce a rotation of 45° in the clockwise sense. The second polarizer P_2 is set at 45° to P_1 so that it will transmit light emerging from the rotator. However, a beam entering from the right will be plane polarized at 45° to the vertical by P_2 and then have its plane rotated by 45° in the clockwise sense by the rotator. It will therefore be incident on P_1 with its plane of polarization at right angles to the plane of transmission and be eliminated. The device thus isolates the components on its left from light incident from the right.

One potential application of magneto-optics currently receiving attention is large capacity computer memories. Such memories must be capable of storing very large amounts of

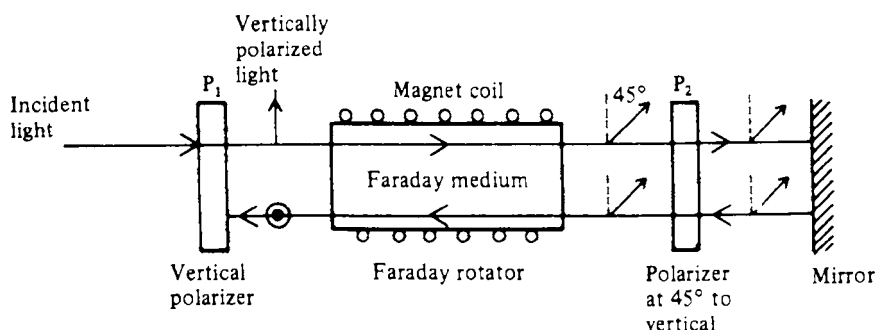


FIG. 3.17 Optical isolator based on the Faraday effect. The reflected ray is shown displaced for clarity.

information in a relatively small area and permit very rapid readout and, preferably, random access. The usual magnetic memories have a number of limitations of size and reading speed. Optical techniques can overcome both these constraints (ref. 3.4).

The magneto-optic memories developed so far are read via the Faraday effect or the magnetic Kerr effect, which relates to the rotation of a beam of plane polarized light reflected from the surface of a material subjected to a magnetic field. In either case a magnetized ferro- or ferrimagnetic material rotates the plane of polarization of laser light incident on it.

Writing may be achieved by heating the memory elements on the storage medium to a temperature above the Curie point using a laser beam. The element is then allowed to cool down in the presence of an external magnetic field thereby acquiring a magnetization in a given direction. Magnetizations of the elements in one direction may represent 'ones', in the opposite direction 'zeros'. To read the information the irradiance of the laser beam is reduced and then directed to the memory elements. The direction of the change in the polarization of the laser beam on passing through or being reflected from the memory elements depends on the directions of magnetization; therefore we can decide if a given element is storing a 'one' or 'zero'.

Systems, incorporating, for example, a 50 mW He-Ne laser with a Pockels modulator and manganese bismuth (MnBi) thin film storage elements, have enabled information to be stored, read and erased at rates in excess of 1 megabits per second.

3.8

Acousto-optic effect

The acousto-optic effect is the change in the refractive index of a medium caused by the mechanical strains accompanying the passage of a surface acoustic (strain) wave along the medium. The strain and hence the refractive index varies periodically with a wavelength Λ equal to that of the acoustic wave. The refractive index changes are caused by the photo-elastic effect which occurs in all materials on the application of a mechanical stress. It can be shown that the change in refractive index is proportional to the square root of the total acoustic power (ref. 3.5).

In general, the relationships between changes of refractive index and mechanical strain,

and between the strain and stress, are rather complicated (ref. 3.6). However, the change in the refractive index can be visualized, as was the case for the electro-optic effect, as a change in the size, shape and orientation of the index ellipsoid. For simplicity let us consider the case of a monochromatic light wave, wavelength λ , travelling in a medium in which an acoustic wave has produced sinusoidal variations of wavelength Λ in the refractive index. The situation is shown in Fig. 3.18, where the solid horizontal lines represent acoustic wave peaks (pressure maxima) and the dashed horizontal lines represent acoustic wave troughs (pressure minima). The portions of the wavefront near to a pressure peak will encounter a higher refractive index and therefore advance with a lower velocity than those portions of the wavefront which encounter pressure minima. The wavefront in the medium therefore soon acquires the wavy appearance shown by the dashed curve in Fig. 3.18. The acoustic wave velocity is very much less than the light wave velocity, so we may ignore it and consider the variation in refractive index to be stationary in the medium.

As elements of the light wave propagate in a direction normal to the local wavefront, almost all the wave elements will suffer a change in direction leading to a redistribution of the light flux, which tends to concentrate near regions of compression. In effect, the acoustic wave sets up a diffraction grating within the medium so that optical energy is diffracted out of the incident beam into the various orders. There are two main cases of interest, namely (a) the *Raman-Nath regime* and (b) the *Bragg regime*.

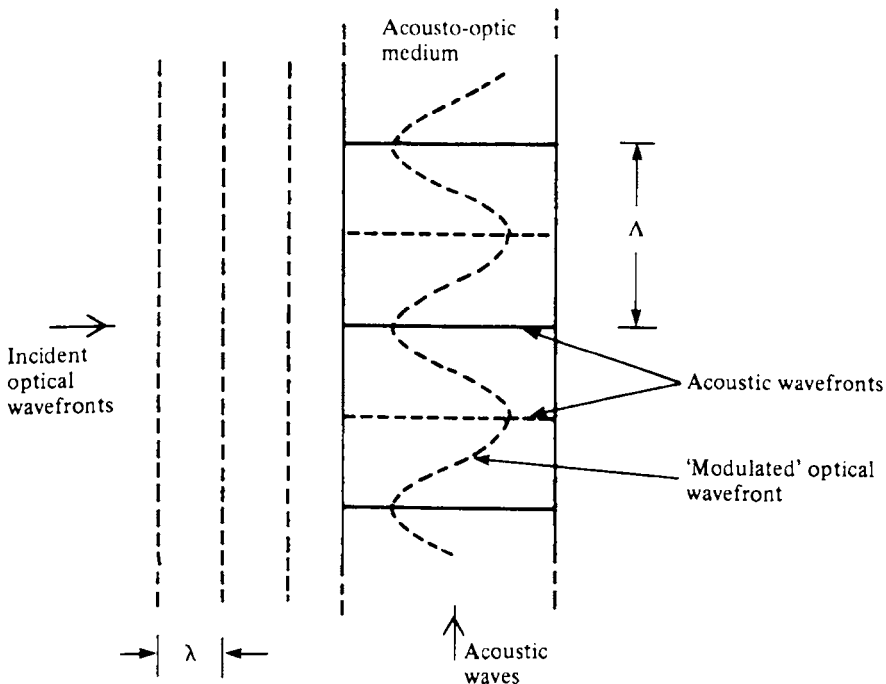


FIG. 3.18 Schematic illustration of acousto-optic modulation. The acoustic waves change the refractive index of the medium in a periodic way so that the plane optical wavefronts take on the 'wavy' appearance shown (very much exaggerated) as they propagate through the medium.

In the Raman–Nath regime the width of the acoustic beam is so small that the diffracted light suffers no further redistribution before leaving the modulator. The light is then diffracted as though from a simple plane grating such that

$$m\lambda = \Lambda \sin \theta_m \quad (3.19)$$

where $m = 0, \pm 1, \pm 2, \dots$ is the order and θ_m is the corresponding angle of diffraction, as illustrated in Fig. 3.19.

The irradiance I of the light in these orders depends on the ‘ruling depth’ of the acoustic grating, which is related to the amplitude of the acoustic grating. This, in turn, is related to the amplitude of the acoustic modulating wave (i.e. the stress produced). The fraction of light removed from the zero-order beam is $\eta = (I_0 - I)/I_0$, where I_0 is the transmitted irradiance in the absence of the acoustic wave. Thus amplitude variations of the acoustic wave are transformed into irradiance variations of the optical beam.

The physical basis of the Bragg regime is that light diffracted from the incident beam is extensively rediffracted before leaving the acoustic field. Under these conditions, the acoustic field acts very much like a ‘thick’ diffraction grating, that is a grating made up of planes rather than lines. The situation is then very similar to that of Bragg diffraction (or ‘reflection’) of X-rays from planes of atoms in a crystal. Consider a plane wavefront incident on the grating planes at an angle of incidence θ_1 as shown in Fig. 3.20(a); significant amounts of light will emerge only in those directions in which constructive interference occurs. The conditions to be satisfied are: (a) light scattered from a given grating plane must arrive in phase at the new wavefront and (b) light scattered from successive grating planes must also

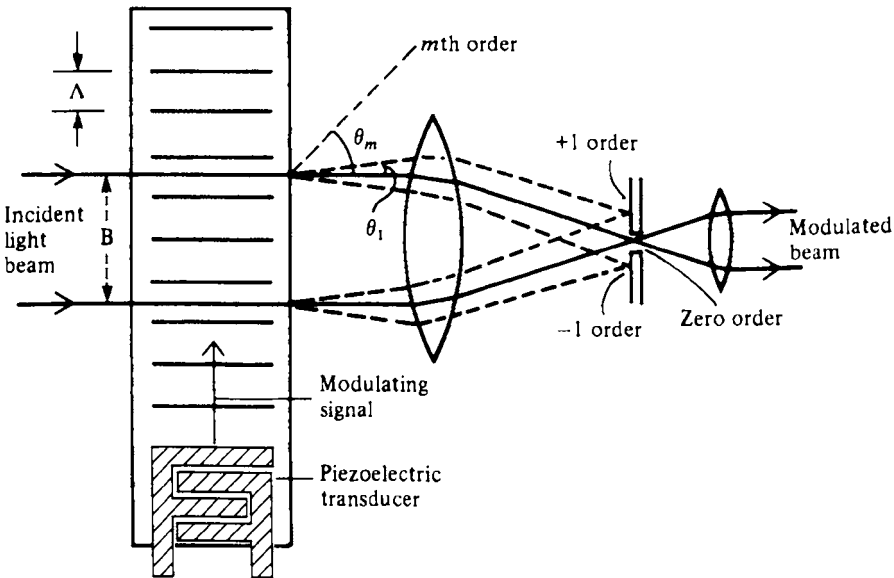


FIG. 3.19 Geometry for Raman–Nath (or transmission-type) acousto-optic diffraction grating modulation. The amount of light diffracted into the orders $m \geq 1$ from the incident beam, and hence the modulation of the transmitted beam, depends on the amplitude of the modulating signal.

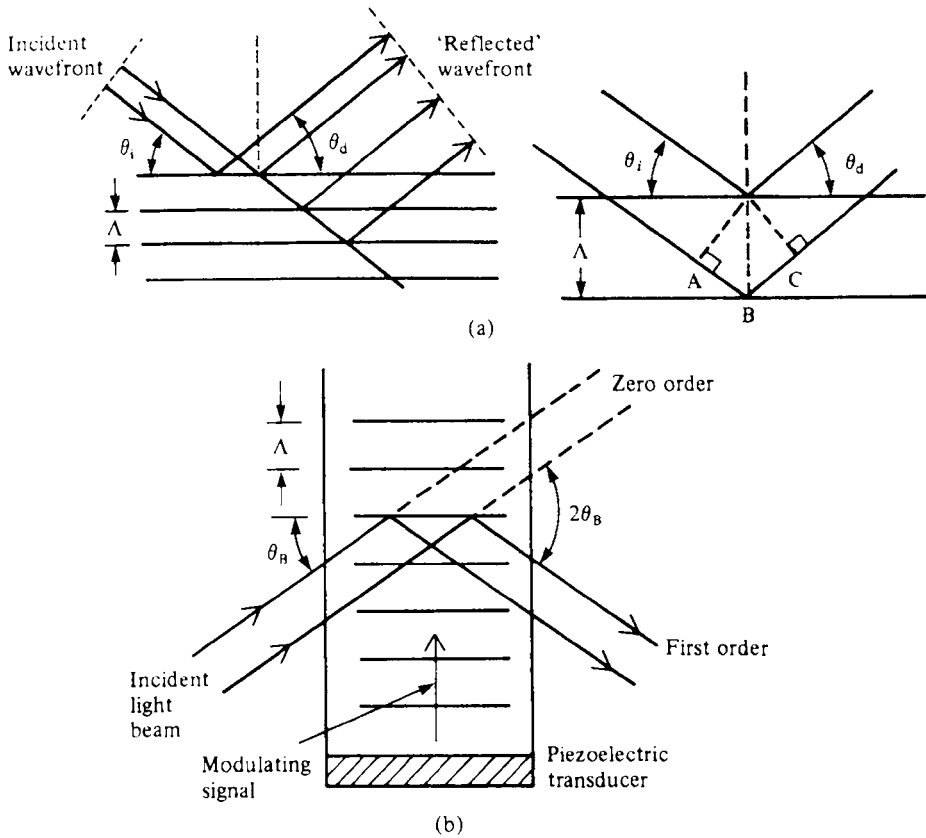


FIG. 3.20 Geometry for Bragg (or reflection-type) acousto-optic diffraction grating modulation: (a) incident rays being scattered from successive layers – for constructive interference the path difference $AB + BC$ must equal an integral number of wavelengths $m\lambda$; (b) the amount of light 'reflected' into the first order depends on the amplitude of the modulating signal (refraction of the light beam at the boundaries of the acousto-optic crystal has been omitted for simplicity).

arrive in phase at the new wavefront, implying that the path difference must be an integral number of wavelengths. The first of these conditions is satisfied when $\theta_d = \theta_i$, where θ_d is the angle of diffraction. The second condition requires that

$$\sin \theta_i + \sin \theta_d = m\lambda/\Lambda$$

with $m = 0, 1, 2, \dots$. The two conditions are simultaneously fulfilled when

$$\sin \theta_i = \sin \theta_d = \frac{m\lambda}{2\Lambda} \quad (3.20)$$

The diffraction is similar to that obtained with a plane grating, but only for special angles of incidence; the angle of incidence must equal the angle of diffraction.

Although in the simplified theory given above strong scattering can take place when m

is equal to any positive integer, a more rigorous treatment, taking into account the fact that scattering is not from discrete planes but from a continuous medium, shows that scattering only takes place when $m = 1$. This is shown in Fig. 3.20(b); the equation for the so-called Bragg angle θ_B then becomes $\sin \theta_B = \lambda/2\Lambda$ (Problem 3.9). The modulation depth $(I_0 - I)/I_0$ (or diffraction efficiency, η) in this case can theoretically equal 100% in contrast to about 34% for the Raman–Nath case. At the Bragg angle η is given by

$$\eta = \sin^2 \phi/2 \quad (3.21)$$

where $\phi = (2\pi/\lambda)(\Delta n L/\cos \theta_B)$, in which Δn is the amplitude of the refractive index fluctuation, L the length of the modulator and θ'_B , the external Bragg angle, is related to θ_B , the internal Bragg angle, by $n \sin \theta_B = \sin \theta'_B$ (ref. 3.7).

The acoustic waves, which create the diffraction grating, are of course moving through the medium and, as a consequence, the diffracted wave behaves as if it had been reflected from a 'mirror' moving with the same velocity as the grating, and therefore appears to originate from a source moving at *twice* the mirror (or grating) velocity. Thus the frequency of the reflected beam is changed by the Doppler effect (see section 5.7) and is given by

$$v' = v_0[1 \pm 2v_x(c/n)]$$

where $\pm v_x$ is the component of the velocity of the acoustic wave along (or away from) the original beam direction and n is the medium refractive index. The frequency shift is thus

$$\Delta v = v' - v_0 = \pm 2v_0 v_x n/c$$

If the light is incident at an angle $\theta_i = \theta_d$ to the acoustic wave as shown in Fig. 3.20(a), then $v_x = v_a \sin \theta_d$, where v_a is the acoustic wave velocity. The frequency shift is then

$$\Delta v = \pm 2v_0 v_a \sin(\theta_d) n/c \quad (3.22)$$

Combining eqs (3.22) and (3.20) and taking $m = 1$ gives a frequency shift of $\pm v_a/\Lambda$ or $\pm f_0$, where f_0 is the acoustic wave frequency. This change in frequency can be used as the basis of a frequency modulator.

The minimum time required to move from a condition where the acoustic wave interacts with the light beam and 'turns off' the undiffracted light to a condition where there is no diffraction is the transit time of the acoustic wave across the optical beam. This is simply, from Fig. 3.19, $t_{\min} = B/v_a$, where B is the optical beam width. Hence the bandwidth of the modulator is limited to about v_a/B . Commercial modulators have bandwidths of up to 50 MHz. This limitation is partly due to the frequency dependence of the acoustic losses of available acousto-optic materials. At the present time only LiNbO_3 , PbMO_4 and TeO_2 appear to have sufficiently low loss to have a reasonable prospect of being operated at appreciably higher frequencies.

Acousto-optic modulators can in general be used for similar applications to electro-optic modulators, though they are not so fast. On the other hand, because the electro-optic effect usually requires voltages in the kilovolt range, the drive circuitry for modulators based on this effect is much more expensive than for acousto-optic modulators, which operate with a few volts.

3.9

Quantum well modulators

We saw in section 2.4.3 that the presence of excitons with binding energies of a few meV in a semiconductor result in the presence of energy levels in the forbidden energy gap, and consequently the absorption of light with photon energies smaller than the bandgap energy E_g . It was also mentioned that in quantum well structures the excitons are more strongly bound, and that the associated absorption is observable at room temperature. If an electric field is applied to a quantum well structure, the electron and hole wavefunctions are separated by being pushed to opposite sides of the well, resulting in a reduction in the absorption, and more importantly a shift of the absorption spectrum to lower energies, that is longer wavelengths. This shift is much larger than in bulk material, and may be ~ 20 meV for electric fields $\sim 10^7$ V m $^{-1}$ in a GaAs/GaAlAs quantum well 10 nm wide (ref 3.8a). This shift, referred to as the quantum confined Stark effect (QCSE), may be used as the basis of an irradiance modulator. At zero field, therefore, light of wavelength slightly longer than the bandgap wavelength λ_g is fully transmitted, while it is strongly attenuated when the field is applied. This is an example of electroabsorption.

An efficient QCSE modulator may comprise a multiple quantum well (MQW) structure, that is a sequence of perhaps 40 or more quantum wells, built into a p-i(MQW)-n diode as shown in Fig. 3.21. The MQWs consist of layers of GaAs 9–10 nm thick separated by layers of AlGaAs of similar thickness. The diode structure is fabricated on a GaAs substrate, which is absorbing at the wavelength of operation so that a window is usually etched into it as shown. Similar devices can be constructed from InGaAs/InAlAs MQWs grown on InP substrates.

In contrast to the vertical geometry shown in Fig. 3.21 it is possible to construct the QCSE

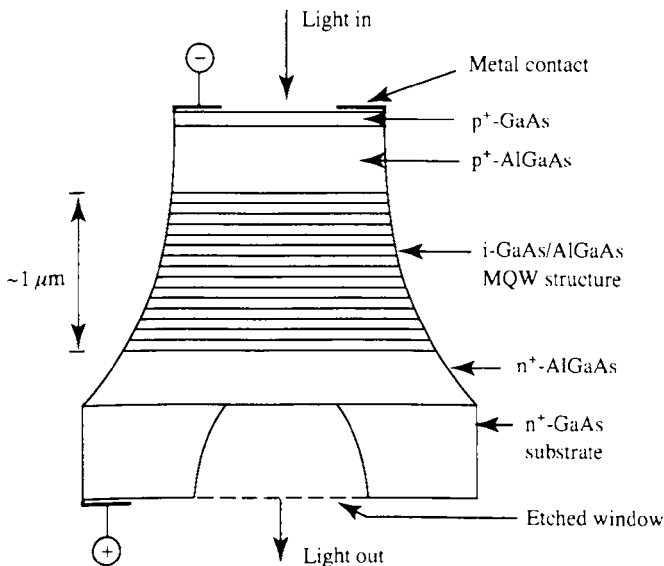


FIG. 3.21 Diagram of a vertical geometry p-i(MQW)-n QCSE modulator.

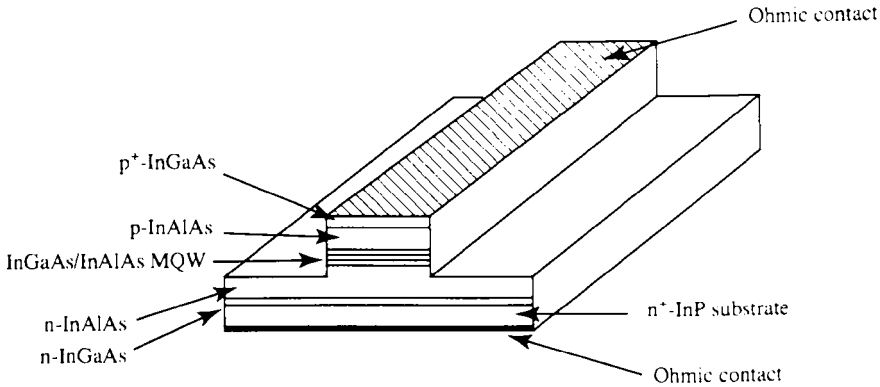


FIG. 3.22 Schematic illustration of a guided wave MQW QCSE modulator.

devices for integrated optical applications (see section 9.4). Here the light to be modulated enters one end of a guided wave structure as shown in Fig. 3.22. The device may be some 200 μm long, in contrast to the 1–2 μm height of the structure in Fig. 3.21, so that comparatively few quantum wells are required to achieve the desired electroabsorption; such devices are also called travelling wave modulators.

The high frequency performance of QCSE modulators depends on what might be regarded as intrinsic factors such as the times that electrons and holes remain in the quantum wells, and extrinsic factors such as the RC time constant as in any junction diode. With care modulation bandwidths greater than 20 GHz can be achieved.

EXAMPLE 3.4 Acousto-optic modulator

Given the following data for a PbMO_4 acousto-optic modulator, we may calculate the Bragg angle, the maximum change in refractive index of the material and the maximum width of the optical beam of wavelength $\lambda = 633 \text{ nm}$ that may be modulated with a bandwidth of 5 MHz.

The modulator length is 50 mm, diffraction efficiency 70%, while the acoustic wavelength is $4.3 \times 10^{-5} \text{ m}$ and the acoustic velocity is 3500 m s^{-1} .

The angle of diffraction (from eq. 3.20) is

$$\theta_B = \sin^{-1} \left(\frac{633 \times 10^{-9}}{2 \times 4.3 \times 10^{-5}} \right) = 7.4 \text{ mrad (or } 0.42^\circ)$$

The value of ϕ is given by eq. (3.21)

$$\phi = 2 \sin^{-1} \sqrt{\eta} = 2 \sin^{-1} \sqrt{0.7}$$

$$\phi = 113.6^\circ$$

Therefore

$$\Delta n = \frac{\phi \lambda \cos \theta_B}{2\pi L} = 1.27 \times 10^{-5}$$

The bandwidth is ν_d/B and hence the maximum optical beam width B is

$$\frac{3500}{5 \times 10^6} = 0.7 \text{ mm}$$

3.10 Non-linear optics

Practical applications of non-linear optical effects have arisen as a direct consequence of the invention of the laser. The very high power densities made available by lasers have enabled several phenomena, which were previously regarded as theoretical curiosities, to be observed and exploited.

The explanation of non-linear effects lies in the way in which a beam of light propagates through a solid. The nuclei and associated electrons of the atoms in the solid form electric dipoles. The electromagnetic radiation interacts with these dipoles causing them to oscillate which, by the classical laws of electromagnetism, results in the dipoles themselves acting as sources of electromagnetic radiation. If the amplitude of vibration is small, the dipoles emit radiation of the same frequency as the incident radiation. As the irradiance of the radiation increases, however, the relationship between irradiance and amplitude of vibration becomes non-linear resulting in the generation of harmonics of the frequency of the radiation emitted by the oscillating dipoles. Thus frequency doubling or second-harmonic generation and indeed higher order frequency effects occur as the incident irradiance is increased. The electric polarization (or dipole moment per unit volume) P can be expressed as a power series expansion in the applied electric field \mathcal{E} by

$$P = \epsilon_0(\chi \mathcal{E} + \chi_2 \mathcal{E}^2 + \chi_3 \mathcal{E}^3 + \dots) \quad (3.23)$$

where χ is the linear susceptibility and χ_2, χ_3, \dots are the non-linear optical coefficients. The resulting (non-linear) relationship between P and \mathcal{E} is shown in Fig. 3.23.

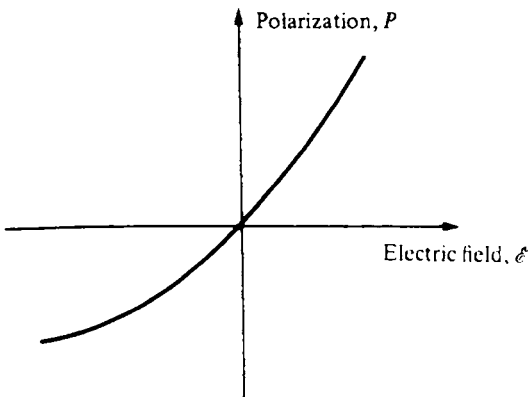


FIG. 3.23 Curve showing electric polarization versus electric field for a non-linear material (i.e. a material lacking a centre of symmetry), assuming χ_2 is negative.

If the applied field is of the form $\mathcal{E} = \mathcal{E}_0 \sin \omega t$ such as produced by an electromagnetic wave, then substitution into eq. (3.23) gives

$$P = \epsilon_0(\chi \mathcal{E}_0 \sin \omega t + \chi_2 \mathcal{E}_0^2 + \chi_3 \mathcal{E}_0^3 \sin^3 \omega t + \dots) \\ = \epsilon_0[\chi \mathcal{E}_0 \sin \omega t + \frac{1}{2} \chi_2 \mathcal{E}_0^2 (1 - \cos 2\omega t) + \dots] \quad (3.24)$$

Equation (3.24) contains a term in 2ω which corresponds to an electromagnetic wave having twice the frequency of the incident wave. The magnitude of the term in 2ω , however, does not approach that of the first term $\epsilon_0 \chi \mathcal{E}_0$ until the electric field is about 10^6 V m^{-1} (which is not entirely negligible in comparison with the internal fields of crystals, i.e. $\mathcal{E}_{\text{int}} \approx 10^{11} \text{ V m}^{-1}$). A field of 10^6 V m^{-1} corresponds, at optical wavelengths, to a power density of about 10^9 W m^{-2} while the electric fields and power density of sunlight are of the order of 100 V m^{-1} and 20 W m^{-2} respectively, so it is not too surprising that the observation of non-linear effects had to await the advent of the laser. (Large non-linear effects have been observed in some semi-conductors with power densities of only about $5 \times 10^4 \text{ W m}^{-2}$ due to free carrier effects.)

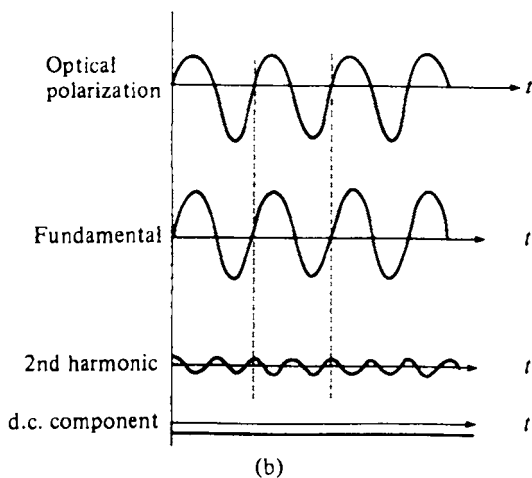
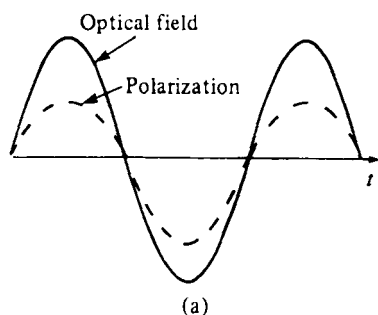


FIG. 3.24 Applied sinusoidal optical (i.e. electrical) field and the resulting polarization for a non-linear material (a) and Fourier analysis of the asymmetrical polarization wave (b) into (i) a fundamental wave oscillating at the same angular frequency (ω) as the wave inducing it, (ii) a second harmonic of twice that frequency (2ω) and (iii) an average (d.c.) negative component.

Harmonic generation is only observed in those solids that do not possess a centre of symmetry. In symmetric materials, an applied electric field produces polarizations of the same magnitude but of opposite sign according to whether the electric field is positive or negative and there is no net polarization. Consequently the coefficients of even powers of \mathcal{E} in eq. (3.24) are zero. In anisotropic media such as quartz, ADP and KDP, however, harmonics are generated as indicated in Fig. 3.24(a) where we see that the symmetrical optical field produces an asymmetrical polarization. A Fourier analysis of the polarization (Fig. 3.24b) shows that it consists of components having frequencies ω and 2ω as well as a d.c. component.

Second-harmonic generation was first observed in 1961 by Franken and his co-workers (ref. 3.9), who focused the 694.3 nm output from a ruby laser onto a quartz crystal as shown in Fig. 3.25 and obtained a very low intensity output at a wavelength of 347.15 nm. In these experiments the conversion efficiency from the lower to the higher frequency was typically $10^{-6}\%$ to $10^{-4}\%$. The reason for this is that wavelength dispersion within the crystal causes the frequency-doubled light to travel at a different speed from that of the fundamental. As the latter is generating the former throughout its passage through the crystal, the two waves periodically get out of phase and destructive interference occurs. The irradiance of the frequency-doubled light thus undergoes fluctuations through the crystal with a periodicity of l_c , which is called the *coherence length* and is typically only a few micrometres.

We can derive an expression for l_c as follows. Let us consider a plane wave propagating through an anisotropic crystal; the fundamental wave has a space-time variation of the form $\exp[i(k_1 z - \omega t)]$, whereas that of the second harmonic is $\exp[i(k_2 z - 2\omega t)]$. The amplitude of the second harmonic as it emerges from the crystal can be found by summing the contributions for the conversion which occurs in each element dz within the crystal, that is

$$\mathcal{E}(2\omega, L) \propto \int_0^L \mathcal{E}^2(\omega, z) dz \quad (3.25)$$

where L is the thickness of the crystal. If we let the time taken for the optical disturbance of frequency 2ω to travel from each point z to L be τ , then we can write relation (3.25) as

$$\mathcal{E}(2\omega, L) \propto \int_0^L \exp\{2i[k_1 z - \omega(t - \tau)]\} dz$$

where

$$\tau = \frac{L - z}{c_{2\omega}} = \frac{(L - z)k_2}{2\omega}$$

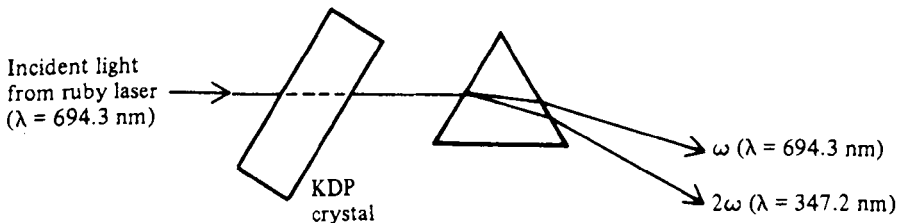


FIG. 3.25 Simplified diagram showing the arrangement for optical frequency doubling. The KDP crystal is mounted in the correct orientation for index matching.

$c_{2\omega}$ being the speed of the second harmonic and k_2 being the corresponding wavevector. Substituting for τ we have

$$\mathcal{E}(2\omega, L) \propto \int_0^L \exp \left\{ 2i \left[\left(k_1 - \frac{k_2}{2} \right) z + \frac{k_2 L}{2} - \omega t \right] \right\} dz$$

Integrating and squaring we find that the irradiance of the second harmonic is proportional to

$$|\mathcal{E}(2\omega, L)|^2 \propto L \left(\frac{\sin[(k_1 - k_2/2)L]}{(k_1 - k_2/2)} \right)^2 \tag{3.26}$$

Relation (3.26) indicates that the irradiance of the second harmonic reaches a maximum after the waves have propagated a distance $L = l_c = \pi/(2k_1 - k_2)$ into the crystal. Thereafter, the energy in the second harmonic is returned to the fundamental wave and after two or indeed any even number of coherence lengths, the irradiance of the second harmonic falls to zero.

This difficulty can be overcome by a technique known as *index or phase matching*. The commonest method uses the birefringent properties of the non-linear medium, which of course must be anisotropic if harmonic generation is to occur at all. As we have illustrated in Fig. 3.26, which shows the normal or index surfaces for n_o^ω and $n_e^{2\omega}$ (section 3.2), it is possible to choose a direction through the crystal such that the velocity of the fundamental (corresponding to the O-ray of frequency ω and refractive index n_o^ω) is the same as that of the

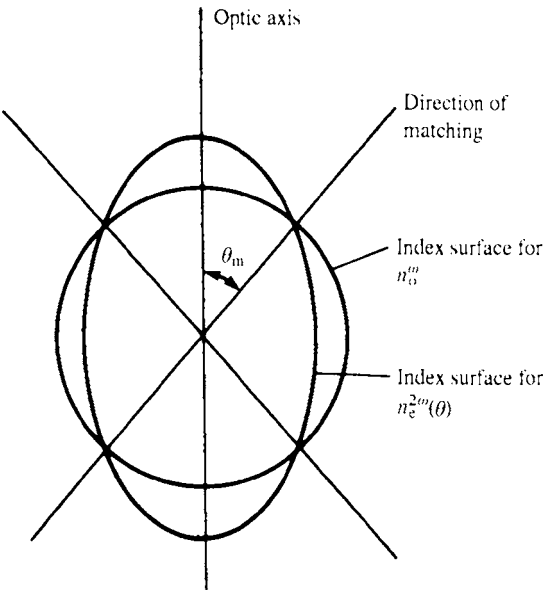


FIG. 3.26 Index matching in a negative uniaxial crystal. The condition $n_o^{2\omega} = n_o^\omega$ is satisfied for propagation at an angle θ_m to the optic axis ($\theta_m \approx 50^\circ$ in KDP). The eccentricities and velocity differences are greatly exaggerated for clarity – similarly the E-ray index surface at frequency ω and the O-ray index surface at frequency 2ω have been omitted.

second harmonic (corresponding to the E-ray of frequency 2ω and refractive index $n_e^{2\omega}(\theta)$); that is, we choose a direction such that $n_o^\omega = n_e^{2\omega}(\theta)$. With this technique, known as index matching, the coherence length increases from a few micrometres to a few centimetres. The conversion efficiency can be increased by orders of magnitude as the fundamental and harmonic waves remain in phase and there is a continuous build-up in the irradiance of the second harmonic. The conversion may be about 20% for a single pass through a KDP crystal a few centimetres long. However, care must be taken in a number of respects to maximize the efficiency. For example, as the refractive indices are temperature dependent it may be necessary to control the crystal temperature. Furthermore, as only one direction of propagation is perfectly index matched, the laser beam divergence must be minimized and similarly lasers with broad linewidths may have lower conversion efficiencies.

EXAMPLE 3.5 Phase matching angle in ADP

Given the following information for ADP determine the angle θ_m . The appropriate refractive indices are

λ (μm)	n_o	n_e
1.06	1.4943	1.4603
0.53	1.5132	1.4712

The phase matching angle is given by (see Problem 3.10)

$$\sin^2 \theta_m = \frac{(n_o^\omega)^{-2} - (n_o^{2\omega})^{-2}}{(n_e^{2\omega})^{-2} - (n_o^{2\omega})^{-2}}$$

Then substituting the numerical values gives

$$\sin^2 \theta_m = \frac{0.4478 - 0.4367}{0.4620 - 0.4367} = 0.44$$

whence $\theta_m = 26^\circ$.

Second-harmonic generation enables us to extend the range of laser wavelengths into the blue and ultraviolet parts of the spectrum, which are not rich in naturally occurring laser lines. One very important laser application which may benefit from second-harmonic generation is laser-induced nuclear fusion, which appears to be more efficient at higher optical frequencies (see section 6.8).

Non-linear processes can be described in terms of a photon model, according to which we can view second-harmonic generation as the annihilation of two photons of angular frequency ω and the simultaneous creation of one photon of frequency 2ω . That is, conservation of energy requires that

$$\hbar\omega + \hbar\omega = \hbar(2\omega) \quad (3.27)$$

while conservation of momentum for the photons similarly requires that

$$\hbar k^{\omega} + \hbar k^{\omega} = \hbar k^{2\omega}$$

or

$$2k^{\omega} = k^{2\omega} \quad (3.28)$$

It is left to the reader to show that this equation is equivalent to the above refractive index matching criterion. The photon model forms a useful basis for the discussion of a related non-linear phenomenon – that of parametric amplification and oscillation.

3.10.1 Parametric oscillation

Second-harmonic generation can be regarded as a special case of sum frequency conversion, whereby power from a ‘pump’ wave at angular frequency ω_3 is transferred to waves at frequencies ω_1 and ω_2 ; that is,

$$\omega_3 \rightleftharpoons \omega_1 + \omega_2 \quad (3.29)$$

The ‘reaction’ represented by relation (3.29) goes to the left in odd-numbered coherence lengths and to the right in even-numbered ones. If the frequency ω_3 alone is applied to a suitable non-linear material such as lithium niobate, then the two smaller frequencies can build up from noise; ω_1 and ω_2 are known as the ‘signal’ and ‘idler’ frequencies respectively. The particular subdivision into ω_1 and ω_2 is determined by the index matching criterion, which conservation of momentum gives as

$$\hbar k_1 + \hbar k_2 = \hbar k_3 \quad (3.30)$$

EXAMPLE 3.6 Coherence length in second-harmonic generation

Given that n^{ω} at $0.8 \mu\text{m}$ is 1.5019 and $n^{2\omega}$ at $0.4 \mu\text{m}$ is 1.4802 in KDP, we can calculate the coherence length as follows: the coherence length is given by $l_c = \pi/(2k_1 - k_2)$, which we can show can be written as

$$l_c = \frac{\lambda_0}{4(n^{\omega} - n^{2\omega})}$$

where λ_0 is the vacuum wavelength at the fundamental frequency. Therefore, using the data provided

$$l_c = \frac{0.8 \times 10^{-6}}{4 \times 0.0217} \approx 10^{-5} \text{ m}$$

It can be seen from relation (3.29) that if ω_3 and ω_1 are fixed, then ω_2 is also fixed ($\omega_2 = |\omega_3 - \omega_1|$). If, however, only frequency ω_3 is fixed, then the other two are free to range over many values; this effect is known as *parametric amplification*. (The term parametric derives from electronic engineering, where so-called parametric amplifiers were developed

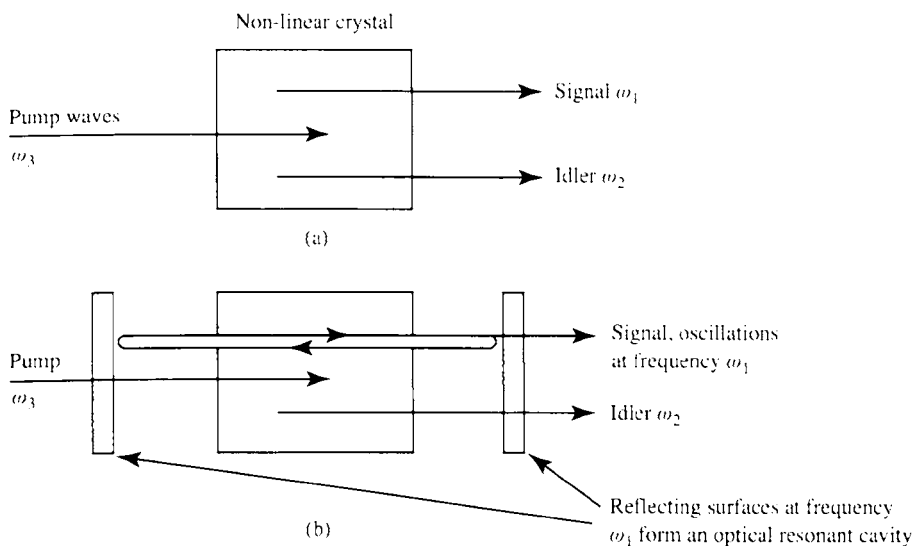


FIG. 3.27 (a) The optical parametric amplifier in which frequencies ω_1 and ω_2 are generated by the pump waves, frequency ω_3 . In (b) the non-linear crystal is placed within an optical cavity resonant at frequency ω_1 , so that oscillations at frequency ω_1 are generated.

in the 1950s. The interaction between the signals at different frequencies was achieved by varying some circuit 'parameter', typically the capacitance.)

If the material producing parametric amplification is placed in an optical resonator (Fig 3.27), such as a Fabry–Perot cavity (see sections 1.23 and 5.5), tuned to the signal, idler or both frequencies, then *parametric oscillations*, which are coherent, will occur – see section 5.10. On the other hand, if energy flows into the wave of frequency ω_3 from the signal and idler waves, the process is called *sum* or *difference frequency generation*, and as mentioned above second-harmonic generation, or frequency doubling, is a special case of sum frequency generation.

The index matching criterion (eq. 3.30) is very severe and the process is usually carried out in an optical cavity (Chapter 5) with mirrors that are highly reflecting at ω_1 or ω_2 but not at ω_3 . Tuning can be achieved simply by varying the index matching conditions through, for example, mechanical or temperature control of the length of the cavity. A schematic diagram of the system used by Giordmaine and Miller (ref. 3.10), who first achieved parametric oscillation in 1965, is shown in Fig. 3.28. In this case, the output was tuned by changing the temperature of the lithium niobate crystal. A temperature change of about 11°C produced output frequencies in the range of 3.1×10^{14} to 2.6×10^{14} Hz (which corresponds to the wavelength range 968–1154 nm). While the conversion efficiency in these experiments was only 1%, efficiencies of about 50% have now been achieved, with coherent wavelengths ranging from the infrared to the ultraviolet being produced.

The additional laser wavelengths generated via parametric processes are useful in extending the range of wavelengths available from the various lasers described in section 5.10. The tunability of the wavelength generated is a particularly useful feature in many applications

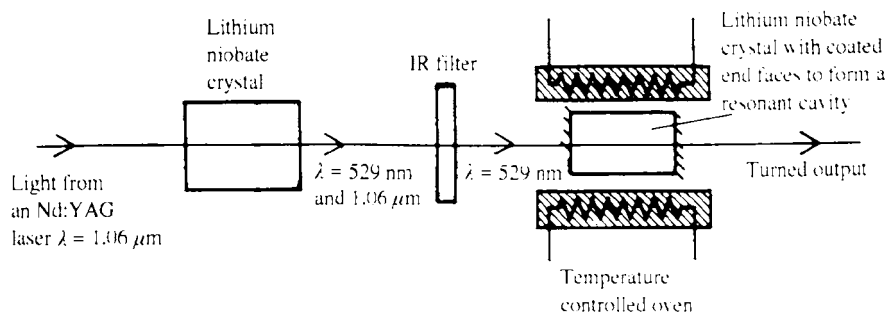


FIG. 3.28 Arrangement of the components in Giordmaine and Miller's observation of parametric oscillation in lithium niobate.

such as photochemistry and the study of the kinetics of ultrafast chemical reactions (see the references on pp. 511–13 of ref. 6.4d).

PROBLEMS

- 3.1 Prove that the angle of inclination of the axes of an ellipse (which represents elliptically polarized light at a fixed position in space) to the horizontal and vertical directions is given by $\frac{1}{2} \tan^{-1}[(\epsilon_0 \epsilon'_0 \cos \phi)/(\epsilon_0'^2 - \epsilon_0^2)]$ (the symbols as defined in section 3.1).
- 3.2 The principal refractive indices for quartz, n_e and n_o , are 1.553 36 and 1.544 25 respectively; calculate the thickness of a quarter-wave plate for sodium D light $\lambda_0 = 589.3$ nm. How may such a plate be made more robust? Calculate the thickness of a calcite $\lambda/2$ plate for which $n_o = 1.658$ and $n_e = 1.486$ at the same wavelength.
- 3.3 Show how a beam of plane polarized light may be considered as consisting of two oppositely directed, circularly polarized beams of light. Hence show that the rotation of the plane of polarization in an optically active medium of thickness d is $(\pi d/\lambda_0)(n_r - n_l)$. The specific rotation of quartz is $29.73^\circ \text{ mm}^{-1}$ at $\lambda_0 = 508.6$ nm; calculate the difference in refractive indices $(n_r - n_l)$.
- 3.4 Using the data given in Table 3.1 calculate some typical half-wave voltages. Why would you expect the half-wave voltage to vary with wavelength?
- 3.5 Calculate the half-wave or switching voltage for a transverse field Pockels modulator using lithium niobate at a wavelength of 632.8 nm where the value of the appropriate electro-optic coefficient is 6.8 pm V^{-1} .
- 3.6 Suggest the design of a double-prism KD*P beam deflector to give a beam deflection of 2° (the voltage applied to the crystal should not exceed 20 kV).
- 3.7 Using the data provided in Table 3.1 calculate the modulation power required for a longitudinal GaAs modulator to give a retardation of π . Take the modulation bandwidth to be 0.1 MHz, the length of the crystal to be 30 mm and the cross-sectional area to be 100 mm^2 . Verify that the same power would be required if the crystal were used in the transverse mode.

- 3.8 Design an optical isolator using zinc sulfide (see Table 3.3); take the permeability of zinc sulfide to be unity and assume that the magnetic field is produced by a solenoid wound directly onto the zinc sulfide at the rate of 5 turns per mm.
- 3.9 Show that, assuming a Bragg acousto-optic modulator is in the form of a parallel-sided block with the acoustic wave travelling perpendicularly to these faces (Fig. 3.20b), the external ray deflection angle is given by $\sin \theta'_B = \lambda_0/2\Lambda$. Hence calculate the beam deflection in a silica modulator, where the acoustic wave velocity is $3.76 \times 10^3 \text{ m s}^{-1}$ and the frequency of the acoustic wave is 50 MHz, for light of $1.06 \mu\text{m}$ wavelength.
- 3.10 Show that the phase matching angle θ_m in a non-linear optical medium is given by

$$\sin^2 \theta_m = \frac{(n_o^\omega)^{-2} - (n_o^{2\omega})^{-2}}{(n_e^\omega)^{-2} - (n_o^{2\omega})^{-2}}$$

(Hint: from Fig. 3.26 it can be seen that θ_m is the angle at which the O-ray velocity at frequency ω equals the E-ray velocity at frequency 2ω , i.e. it is the angle at which the ellipse and circle in the figure intersect; the value of extraordinary refractive index as a function of θ is given in Appendix 2.)

REFERENCES

- 3.1 (a) G. R. Fowles, *Introduction to Modern Optics* (2nd edn), Holt, Rinehart & Winston, New York, 1975, Chapter 6.
(b) E. Hecht, *Optics* (2nd edn), Addison-Wesley, Reading, MA, 1987, Chapter 8.
- 3.2 (a) See ref. 3.1a, Chapter 9.
(b) A. Yariv, *Optical Electronics* (4th edn), Holt-Saunders, Japan, 1991.
- 3.3 (a) J. F. Nye, *Physical Properties of Crystals*, Oxford University Press, Oxford, 1957, Chapter 13.
(b) I. P. Kaminow, *An Introduction to Electro-optic Devices*, Academic Press, New York, 1974.
(c) R. Guenther, *Modern Optics*, John Wiley, Toronto, 1990.
- 3.4 (a) D. Chen, 'Magnetic materials for optical recording', *Appl. Opt.*, **13**, 767, 1974.
(b) D. Chen and J. D. Zook, 'An overview of optical data storage technology', *Proc. IEEE*, **63**, 1207, 1975.
- 3.5 D. A. Pinnow, *IEEE J. Quantum Electron.*, **QE-6**, 223, 1970.
- 3.6 (a) A. Yariv, *Quantum Electronics*, John Wiley, New York, 1975, Sections 14.8–11.
(b) See ref. 3.2b, Chapter 12.
- 3.7 L. Levi, *Applied Optics*, Vol. II, John Wiley, New York, 1980, Chapter 14.
- 3.8 (a) J. Singh *et al.*, 'System requirements and feasibility studies of optical modulators based on GaAs/AlGaAs multiquantum well structures for optical processing', *J. Lightwave Technol.*, **6**, 818, 1988.

- (b) P. Bhattacharya, *Semiconductor Optoelectronic Devices* (2nd edn), Prentice Hall, Englewood Cliffs, NJ, 1997, sections 3.4 and 11.4.
- 3.9 P. A. Franken, A. E. Hill, C. W. Peters and G. Weinreich, 'Generation of optical harmonics', *Phys. Rev. Lett.*, **7**, 118, 1961.
- 3.10 J. A. Giordmaine and R. C. Miller, 'Tunable optical parametric oscillation in LiNbO_3 at optical frequencies', *Phys. Rev. Lett.*, **14**, 973, 1965.

Display devices

We may divide display devices into two broad categories: (a) those that emit their own radiation (active devices) and (b) those that in some way modulate incident radiation to provide the display information (passive devices). However, before discussing the devices themselves, we consider the circumstances under which matter can be induced to emit radiation.

4.1 Luminescence

Luminescence is the general term used to describe the emission of radiation from a solid when it is supplied with some form of energy. We may distinguish between the various types of luminescence by the method of excitation. For example:

<i>Photoluminescence</i>	excitation arises from the absorption of photons
<i>Cathodoluminescence</i>	excitation is by bombardment with a beam of electrons
<i>Electroluminescence</i>	excitation results from the application of an electric field (which may be either a.c. or d.c.)

Whatever the form of energy input to the luminescing material, the final stage in the process is an electronic transition between two energy levels, E_1 and E_2 ($E_2 > E_1$), with the emission of radiation of wavelength λ_0 where (see section 1.3)

$$\frac{hc}{\lambda_0} = E_2 - E_1 \quad (4.1)$$

Invariably E_1 and E_2 are part of two *groups* of energy levels, so that, instead of a single emission wavelength, a *band* of wavelengths is usually observed.

When the excitation mechanism is switched off, we would expect the luminescence to persist for a time equal to the lifetime of the transition between the two energy levels E_1 and E_2 . When this is so, we speak of *fluorescence*. Often, however, the luminescence persists for much longer than expected, a phenomenon called *phosphorescence*. Phosphorescence is often attributable to the presence of metastable (or very long lifetime) states with energies less than E_2 . Electrons can fall into these states and remain trapped there until thermal excitation releases them some time later. Materials exhibiting phosphorescence are known as *phosphors*. Generally speaking, phosphor materials depend for their action on the presence within the material of impurity ions called *activators*. These replace certain of the host ions in the crystal

lattice. Unless the charge on the activator ion is identical with that of the host ion it replaces, the charge balance will be upset and few will be able to enter the lattice. Improved solubility of the activator in these circumstances may result from the introduction of further impurity atoms with different ionic charge. These are known as *co-activators*.

We may distinguish between two main types of energy level system. In the first the energy levels are those of the activator ion itself, whilst in the second they are those of the host lattice modified by the presence of the activator ions. We refer to these two types as 'characteristic' and 'non-characteristic' respectively.

In characteristic luminescence, the excitation energy is usually transferred rapidly (i.e. in a time very much less than 10^{-8} s) to the activator ion. The persistence of the luminescence is then entirely due to the lifetime of the excited state level of the activator. It should be noted that, whilst for atomic electric dipole transitions this is of the order of 10^{-8} s, it can be much longer if such transitions are forbidden (see ref. 4.1). Hence fluorescence cannot be unambiguously associated with characteristic luminescence.

In non-characteristic luminescent materials, both activators and co-activators are usually present. These create acceptor and donor energy levels in the material (see section 2.4.2), although in phosphors these levels are usually referred to as hole and electron traps respectively. Energy absorption within the solid creates excess electron-hole pairs and, as the hole trapping probability is usually much greater than the electron trapping probability, most of the excess holes quickly become trapped. Any electron that then finds itself in the vicinity of a trapped hole can recombine with it and generate luminescence. As the electrons migrate through the crystal, however, they themselves are subject to trapping. The electron traps could, of course, act as recombination centres were there appreciable numbers of free holes present, but the difference in trapping probability prevents this. Instead an electron may remain in its trap for some time before subsequently being released by thermal excitation. It may then go on to be retrapped or to recombine with a trapped hole. This process is illustrated in Fig. 4.1.

The time that an electron spends in a trap depends on the depth of the trap below the conduction band ($E_c - E_d$) and also the temperature T . It is generally found that the probability of escape per unit time can be written in the form $Q \exp\{-(E_c - E_d)/kT\}$ where Q is

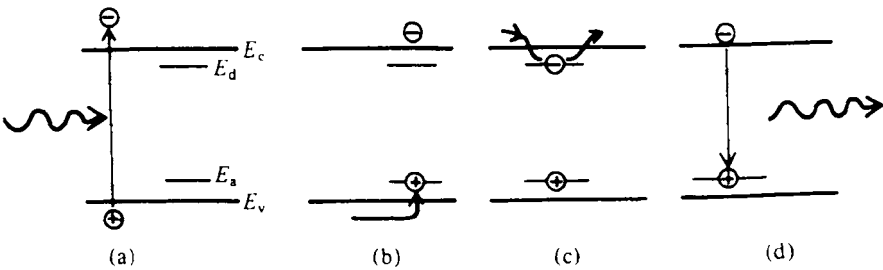


FIG. 4.1 Electron-hole generation and recombination processes in non-characteristic luminescent materials. Electron-hole pairs are generated by photon absorption (a), and the holes are quickly trapped at acceptor sites (b). Electrons may then recombine with these trapped holes, thereby giving rise to luminescent emission (d). However, before such a recombination can take place, the electron itself may spend some time trapped at a donor site (c).

a constant approximately equal to 10^8 s^{-1} (ref. 4.2). Thus with fairly 'deep' traps (i.e. large $E_c - E_d$) and at low temperatures the time spent in a trap may be comparatively large, which will cause a long persistence of the luminescence after the cessation of excitation. In some cases this may amount to hours or even days.

EXAMPLE 4.1 Luminescence lifetime due to traps

If we take a value of 0.4 eV for $E_c - E_d$, then the probability of escape per second of a trapped electron at room temperature (where $kT = 0.025 \text{ eV}$) is of the order of $10^8 \exp(-0.4/0.025) \approx 10 \text{ s}^{-1}$.

Hence we would expect a luminescence lifetime of about 0.1 s.

We turn now to a more detailed discussion of photoluminescence, cathodoluminescence and electroluminescence.

4.2 Photoluminescence

As we have seen, in photoluminescence energy is transferred to the crystal by the absorption of a photon. In characteristic luminescent materials the activator ion itself absorbs the photon directly. It might be expected, therefore, that since the same energy levels are involved in absorption as in emission then the wavelengths for absorption and emission would be identical. In fact it is found that the peak emission wavelength is invariably shifted towards the red end of the spectrum compared with the peak of the absorption spectrum. This phenomenon is known as the Stokes shift, and it may be understood by taking account of the effect of the vibrations of the surrounding crystal lattice on the energy levels of the activator ions. The

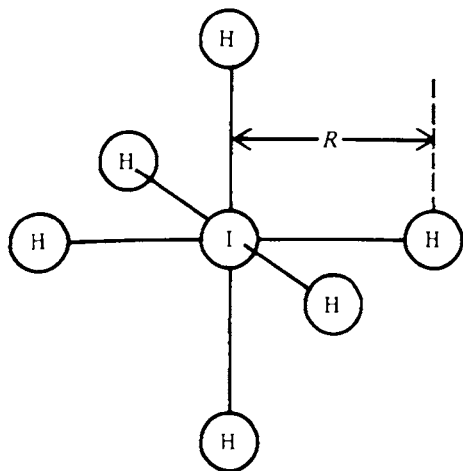


FIG. 4.2 Assumed impurity site structure in a characteristic luminescent material. The impurity ion (I) is surrounded by six host ions (H) each at a distance R from it.

latter are often positively charged, typical examples being Cr^{3+} and Mn^{2+} , although the exact charge state depends on the host. For the sake of argument, we assume here that each activator ion is positively charged and is surrounded by six equidistant negatively charged ions at a distance R from the activator as shown in Fig. 4.2. We further assume that in the most important vibrational mode of the group, the activator ion remains at rest whilst the six surrounding negative ions all vibrate radially and in phase. Because of electrostatic interactions, the positions of the energy levels of the activator ion will depend on the value of R . A schematic diagram illustrating this variation for two energy levels is shown in Fig. 4.3. The most important feature of this diagram is that the minima of the two curves do not occur at the same values of R . This is perhaps not too surprising, since the equilibrium distribution of charge round the activator ion will be different when it is in each of the two states.

Consider now the absorption of a photon when the activator is in its ground state, where the most probable value for R is R_0 (the position of minimum energy). Photon absorption is a very rapid process and takes place virtually instantaneously as far as the vibration of the sur-

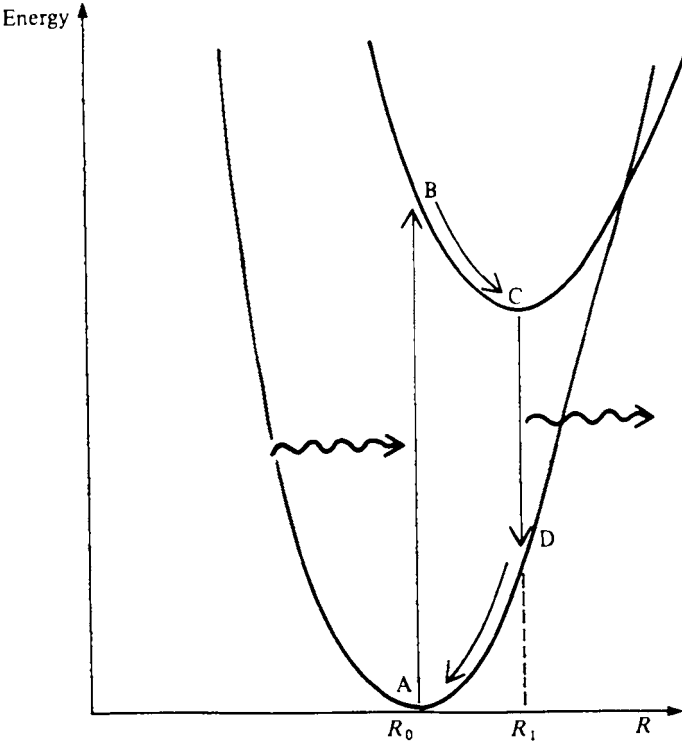


FIG. 4.3 Schematic diagram of the variation in the energy of two electron energy levels of an impurity ion in a characteristic luminescent material as a function of the nearest neighbour ion separation R . If the ion is initially at the point A on the diagram (where $R = R_0$), then photon absorption can take place, and the energy of the ion will change to that of the point B. The surrounding ions then relax to a new equilibrium position (R_1), and the impurity ion moves from B to C, losing energy by phonon emission. The impurity ion may then make a transition to the point D and emit a photon. Once at D the surrounding ions relax back to R_0 and the impurity ion returns to the point A, again losing energy by phonon emission.

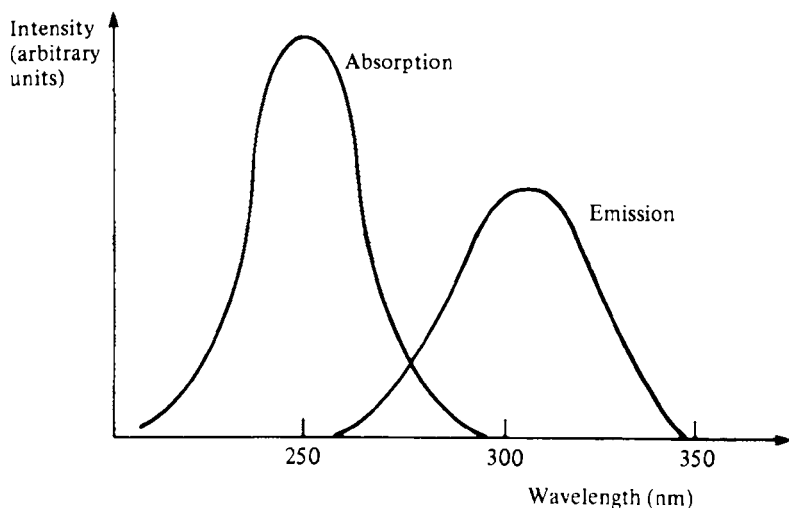


FIG. 4.4 Absorption and emission spectra for thallium-activated potassium chloride (KCl:Tl) at room temperature. The emission peak occurs at a higher wavelength than that of the absorption curve. This is an example of the Stokes shift.

rounding ions is concerned. On the energy level diagram, therefore, the process may be represented by a vertical transition (i.e. with R remaining constant). Immediately after the transition to the excited state the surrounding ions will not be at their equilibrium positions for this state and subsequently they will relax to their new equilibrium positions (at $R = R_1$ in Fig. 4.3). When a downward transition takes place, R again remains constant and an inspection of Fig. 4.3 shows that the emitted photon will then have less energy than that of the absorbed photon.

We have of course been talking about the *most probable* transition. The ions surrounding the activator will always be in a state of oscillation, and hence at the instant of absorption the value of R may well differ from R_0 . The same will be true for the excited state. Thus instead of a single absorption (or emission) wavelength a *band* of absorption (or emission) wavelengths is seen. Further, since the amplitude of the oscillation will increase with increasing temperature, we would expect the width of the absorption and emission bands also to increase with increasing temperature. Figure 4.4 shows a curve of absorption and emission in KCl: Tl and illustrates the Stokes shift seen in this material at room temperature.

The Stokes shift finds commercial application in fluorescent lamps, in which an electrical discharge is passed through a mixture of argon and mercury vapour. The emitted radiation has a bluish colour and an appreciable amount of the radiant energy is in the ultraviolet. If the walls of the discharge tube are coated with a suitable luminescing material, the ultraviolet radiation may be converted to useful visible radiation, thereby increasing the luminous efficiency of the lamp.

4.3

Cathodoluminescence

In cathodoluminescence, although the emission processes are the same as those outlined above for photoluminescence, the excitation mechanisms are somewhat more involved. When

a beam of energetic electrons (say with energy greater than 1 keV) strikes a solid, a fraction (about 10%) is backscattered. The remaining electrons penetrate into the solid where they rapidly lose energy, mainly by causing bound electrons to be ejected from their parent ions. These secondary electrons in turn may generate further secondary electrons, provided they have sufficient energy. The final stages in the energy loss process consist of the excitation of electrons from states at the top of the valence band (with energy E_v) to those at the bottom of the conduction band (with energy E_c). Energy conservation considerations alone would dictate that the exciting electrons must at this stage have an energy of at least $E_c - E_v = E_g$ above E_c if they are to create electron-hole pairs and still remain in the conduction band. In fact the additional constraint of momentum conservation requires a somewhat higher minimum energy of approximately $E_c + 3E_g/2$ (ref. 4.3). When electrons have energies between this value and E_c , they can only lose energy by exciting lattice vibrations (phonons). In addition, of course, even when electrons have higher energies than $E_c + 3E_g/2$, energy may still be wasted in phonon generation. It has been found empirically for a range of semiconductor materials that the total number of electron-hole pairs generated may be written as $E_B/\beta E_g$, where E_B is the total electron beam energy and $\beta \approx 3$. This inefficiency in the generation of electron-hole pairs is a major factor in causing cathodoluminescence to be a considerably less efficient process than photoluminescence.

In non-characteristic materials, electron-hole recombination and luminescent emission then take place as for photoluminescence. In characteristic materials, on the other hand, it is thought that the next step is the formation of excitons (bound electron-hole pairs, see section 2.4.3). These migrate through the lattice and may subsequently transfer their recombination energy to the activator ions.

As the primary electrons rapidly lose energy they penetrate only a little way into the solid they are exciting. It has been found experimentally that the penetration depth, or range, R_e of an electron beam of energy E_B is given by (ref. 4.4)

$$R_e = KE_B^b \quad (4.2)$$

where the parameters K and b depend on the material. For ZnS, for example, the range is in micrometres when $K = 1.2 \times 10^{-4}$ and $b = 1.75$. Thus a 10 keV electron beam has a range in ZnS of $1.2 \times 10^{-4} \times 10^{1.75}$ or 0.7 μm .

It is often found that cathodoluminescent efficiency increases with increasing beam voltage. This may be attributed mainly to the fact that at low beam voltages most of the electron-hole pairs are generated close to the surface of the luminescent material, where there is often a relatively high concentration of non-luminescent recombination centres. (Another instance of a similar deleterious surface effect is the reduction in efficiency of photoconductive detectors with decreasing wavelength, as discussed in section 7.2.7.)

4.4 Cathode ray tube

The relative ease with which a beam of electrons can be directed and focused led to the early development and continued use of the cathode ray tube (CRT) as an important tool for the analysis of rapidly varying electrical signals as well as providing a versatile optical display

device. Only a very brief discussion of the CRT can be attempted here and the reader is referred to ref. 4.5 for more details. Figure 4.5 shows the basic construction. Electrons are generated by thermionic emission (see section 2.6) by heating a specially impregnated cathode surface (usually based on oxides of barium and strontium) and then focused onto the viewing screen by a series of metal electrodes held at various potentials. A grid for the control of the electron flow is usually also included. The whole assembly is known as an 'electron gun'. The electron beam is scanned across the viewing screen in a series of lines; when one line scan is completed the beam is rapidly switched to the start of the line below. Beam deflection is controlled by electrostatic or electromagnetic fields acting at right angles to the beam direction. Electrostatic deflection enables the highest beam deflection rates to be achieved, whilst electromagnetic deflection enables higher beam accelerating potentials to be employed, which results in a smaller spot size and higher screen brightness. When the beam strikes the viewing screen, radiation is generated by cathodoluminescence. The screen consists of a thin layer of small (dimensions $\approx 5 \mu\text{m}$) phosphor granules, with a layer of aluminium ($\approx 0.1 \mu\text{m}$ thick) evaporated onto the gun side (Fig. 4.6). This layer serves two purposes: first it prevents charge build-up on the phosphor granules (which generally have low conductivities) and secondly it helps to reflect light emitted in a direction away from the observer back towards him or her.

The thicknesses of both the aluminium and phosphor layers are fairly critical. If the aluminium is too thick, an appreciable fraction of the electron beam energy will be absorbed within it, whilst if it is too thin its reflectivity will be poor. If the phosphor layer is too thick, scattering and absorption reduce the light output, whilst too thin a layer, on the other hand, can result in incomplete coverage of the screen area.

For normal display operations (e.g. television) the beam is scanned line by line over the

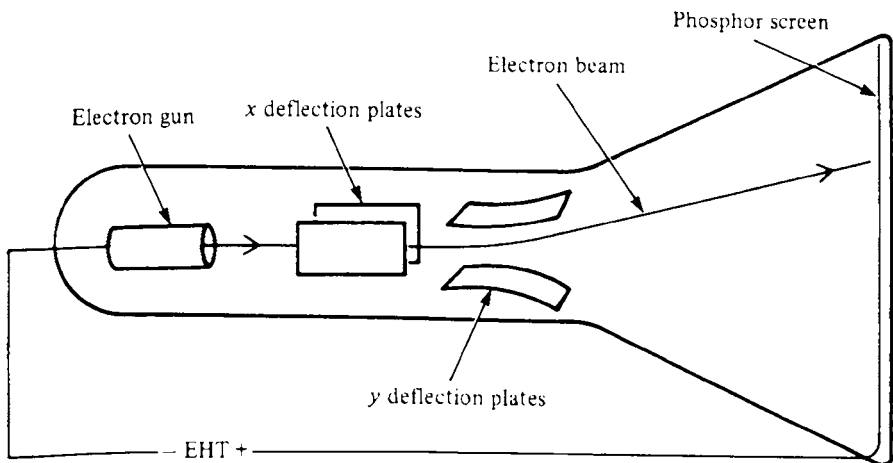


FIG. 4.5 Schematic diagram of a CRT. An electron beam originating at an electron gun passes through an electrostatic deflection system (which uses two sets of plates at right angles, one for the 'x' deflection and the other for the 'y' deflection) and then falls onto a phosphor screen. Details of the electron gun and its associated focusing system are not shown.

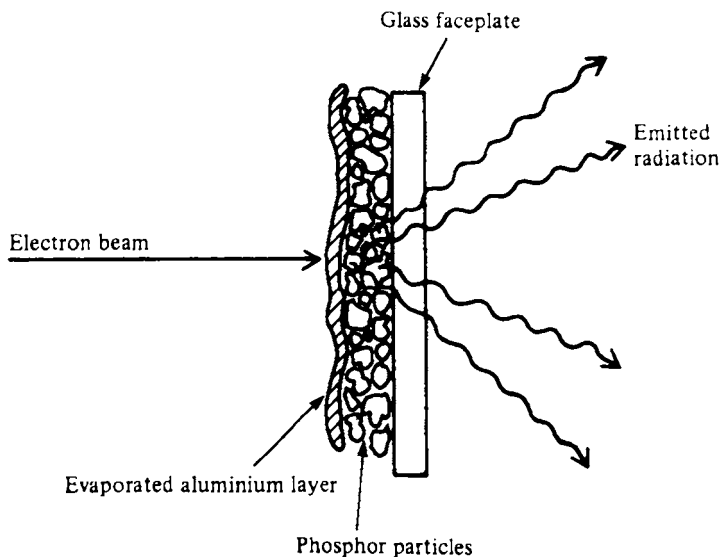


FIG. 4.6 Cross-section of a CRT screen. A layer of phosphor is sandwiched between a glass faceplate and an evaporated aluminium layer. High energy electrons penetrate the aluminium and excite cathodoluminescence in the phosphor particles. The aluminium layer reduces charge build-up and helps to reflect light back out through the faceplate.

viewing area. In video applications the display consists of some 625 lines in Europe and 525 in North America. To avoid an image that 'flickers' the picture must be renewed at a rate greater than about 45 Hz. However, it is possible to avoid having to renew the entire picture at this rate by using a raster scan that splits the picture up into two interlaced halves. Thus if a complete picture scan takes t_i seconds then we may arrange that during the first $t_i/2$ seconds lines 1, 3, 5, 7, etc., are scanned, whilst during the second $t_i/2$ seconds lines 2, 4, 6, 8, etc., are scanned. Because the two images are effectively superimposed, the eye treats the picture repetition rate as if it were $2/t_i$ Hz rather than $1/t_i$ Hz. This reduction in the rate at which picture information is required before flicker becomes troublesome is very useful because it halves the transmission frequency bandwidth that would otherwise be required. In Europe the entire picture is scanned in $1/25$ s, whilst in North America this time is $1/30$ s. Varying light irradiances are obtained by varying the beam current. Ideally, the phosphor used should have a luminescent decay time shorter than the picture cycle time, otherwise streaking effects due to image persistence are obtained. CRT displays can be made sufficiently bright for them to be visible under nearly all ambient lighting conditions. The brightness limits are usually reached when the phosphor screen rapidly deteriorates under high beam currents.

Colour displays for home video viewing are obtained using the 'shadowmask' principle, in which three electron guns are used that are slightly inclined to each other so that their beams coincide at the plane of the shadowmask. The latter is a metal screen, with holes in it, placed just in front of the phosphor screen. Having passed through one of the holes in the

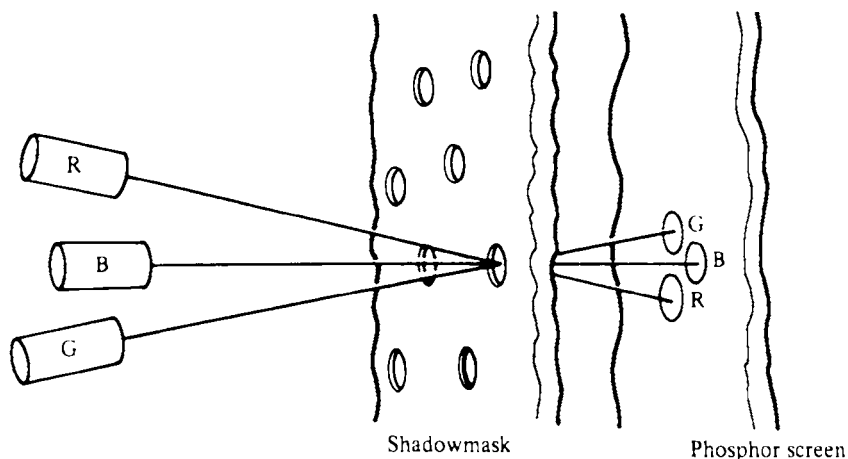


FIG. 4.7 Use of the shadowmask for obtaining colour displays. Three separate guns are used; these are inclined slightly to each other so that their beams will all pass through a single hole in the shadowmask. After passing through the hole, the beams diverge so that each falls on one of the three circular areas composed of phosphors, each of which emits one of the three primary colours (blue, green and red).

shadowmask the three beams diverge, and on striking the phosphor screen are again physically distinct. The phosphor screen consists of groups of three phosphor dots, placed so that when the three beams pass through a hole in the shadowmask they each hit a different dot. Figure 4.7 illustrates the basic geometry. Each of the three phosphor dots emits one of the primary colours (i.e. blue, green and red), so that any desired colour can be generated by varying the relative excitation intensities. Some of the more commonly used phosphors are zinc sulfide doped with silver: ZnS:Ag (blue); zinc cadmium sulfide doped with copper: $\text{Zn}_x\text{Cd}_{1-x}\text{S:Cu}$ (green); and yttrium oxysulfide doped with europium and terbium: $\text{Y}_2\text{O}_2\text{S:Eu,Tb}$ (red). The first two are non-characteristic phosphor materials whilst the last is a characteristic material. There is obviously some loss in resolution capability over a monochrome display since the coloured dots are physically displaced on the screen. Furthermore, the alignment of the shadowmask with the guns and the phosphor screen is critical, and can be spoilt by fairly harsh environmental conditions such as stray magnetic fields. Several modifications have been made to the basic shadowmask principle to try and rectify some of these disadvantages; the interested reader is referred to ref. 4.5 for further information.

To render the screen more easily visible in the presence of ambient radiation the faceplate is usually tinted a shade of grey, with a transmission of about 50%. This increases the contrast between the light emitted from the phosphor particles (which traverses the faceplate once) and the ambient light reflected from the back of the phosphor (which traverses the faceplate twice). A recent proposal has been to put an appropriately coloured 'microfilter' between each area of phosphor and the faceplate. The filters allow most of the radiation emitted by the area of phosphor to pass through, whilst absorbing much of the incident 'white' ambient light. The faceplate can now be made of clear glass, so that the screens are almost 50% brighter (or alternatively achieve the same brightness with the use of less power).

4.5

Electroluminescence

We are concerned here with what might be termed 'classical electroluminescence' as opposed to 'injection electroluminescence', which uses fabricated p-n junctions and which will be dealt with in section 4.6. Four main types of device may be distinguished, depending on the type of drive (a.c. or d.c.) and the character of the active layer (powder or thin film). The first electroluminescent device to be extensively studied was the a.c. powder device, proposed in 1936. In this a phosphor powder (usually ZnS:Cu) is suspended in a transparent insulating binding medium of high dielectric constant and is sandwiched between two electrodes (one of which is transparent) as shown in Fig. 4.8(a). Usually there is no complete conducting path between the electrodes, so that d.c. excitation is not possible. When an alternating voltage, $V_0 \cos(2\pi ft)$, is applied across the cell, however, light is emitted in the form of short bursts which last about 10^{-3} s and occur once every half cycle. It is found that the integrated light output power P can be written in the form (ref. 4.6)

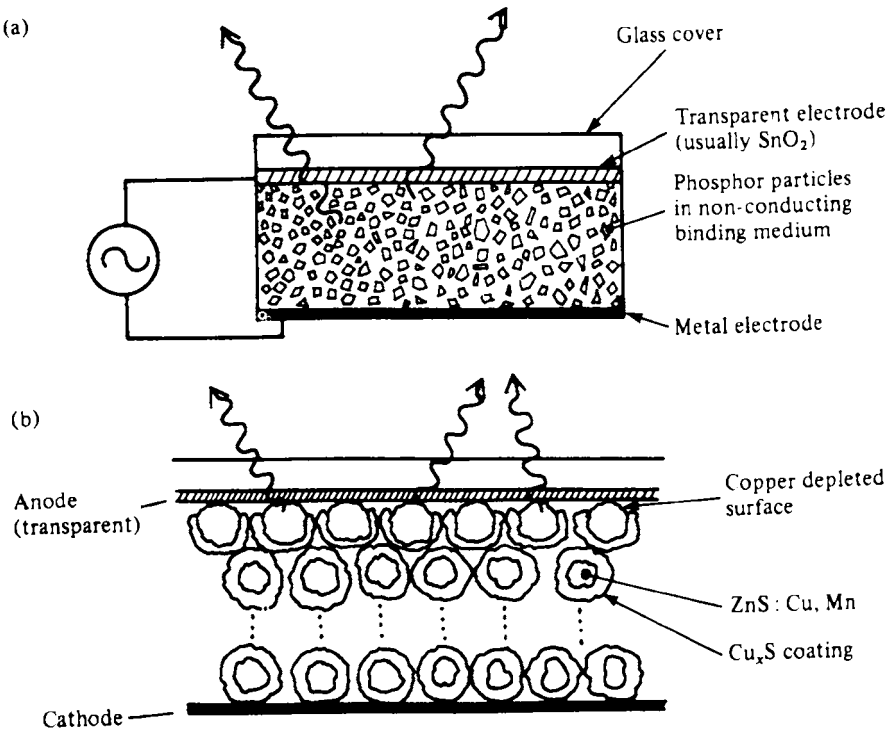


FIG. 4.8 Construction of an a.c. electroluminescent device (a). Phosphor particles are suspended within a transparent insulating medium and sandwiched between two electrodes, one of which is transparent. When an alternating voltage is applied to the electrodes the phosphor particles emit light. (b) The construction of a d.c. electroluminescent device. The phosphor particles have a coating of Cu_xS. This coating is removed from the anode side of the particles in contact with the anode by the application of an initial high current pulse. Under normal conditions, light is emitted only from the Cu_xS depleted particles.

$$P = P_0(f) \exp \left[- \left(\frac{V_1}{V_0} \right)^{1/2} \right] \quad (4.3)$$

where V_1 is a constant and $P_0(f)$ is a function of frequency. The strongest emission from within the phosphor grain is found to take place from the side temporarily facing the cathode. Several possible emission mechanisms have been proposed (see ref. 4.6 for further details); it is generally agreed, however, that there will be a high electric field within the phosphor particle. It is then possible that this field is sufficiently strong to enable electrons from occupied acceptor levels to 'tunnel' to states of the same energy in the conduction band, as illustrated in Fig. 4.9(a). (Quantum mechanical tunnelling was briefly mentioned in section 2 8.5.) Other electrons in the conduction band are then able to fall into these vacated levels and emit radiation (Fig. 4.9b).

Another possibility is that an electron moving in the electric field may acquire sufficient energy to enable it to excite an electron from the valence band to the conduction band. The resulting hole quickly becomes trapped at an impurity acceptor site, thereby effectively emptying it of an electron. An electron in the conduction band can then make a radiative transition by falling into the empty acceptor level. The sequence of events is illustrated in Fig. 4.10. In phosphors containing manganese, there is evidence that the Mn^{2+} ions themselves may be directly excited by the high energy electrons, radiation being emitted when the ion subsequently undergoes de-excitation. A.C. powder devices usually require several hundred volts (r.m.s.) to drive them. They exhibit luminances of about 40 nits (for a definition of the nit, see section 4.8), have power efficiencies (i.e. the ratio of optical power out to electrical power in) of about 1% and lifetimes of about 1000 hours. By using different phosphor powders, red, green, yellow and blue displays are possible.

A more recent development is the d.c. powder display (Fig. 4.8b). These have a structure

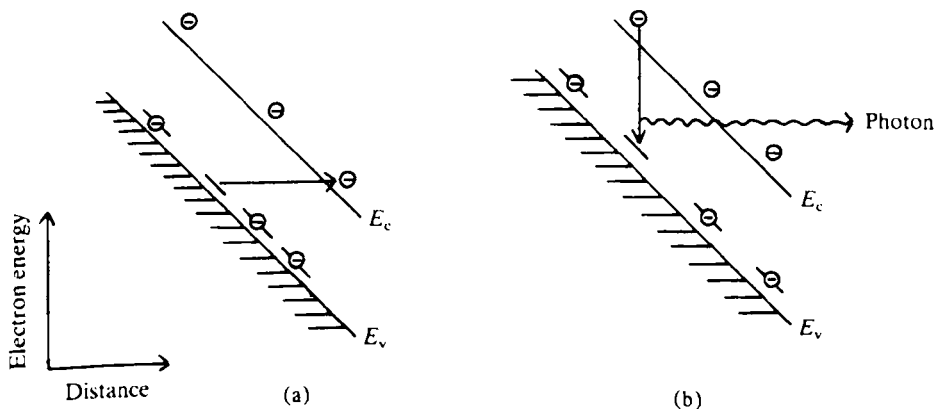


FIG. 4.9 Possible mechanism for electroluminescence emission involving quantum mechanical tunnelling. In (a) an electron in an acceptor state 'tunnels' through the forbidden gap region into states of the same energy. It is only able to do this if there is a considerable electric field present, thus causing the energy bands to be tilted. An electron in the conduction band may now fall into the vacated level resulting in radiative emission (b).

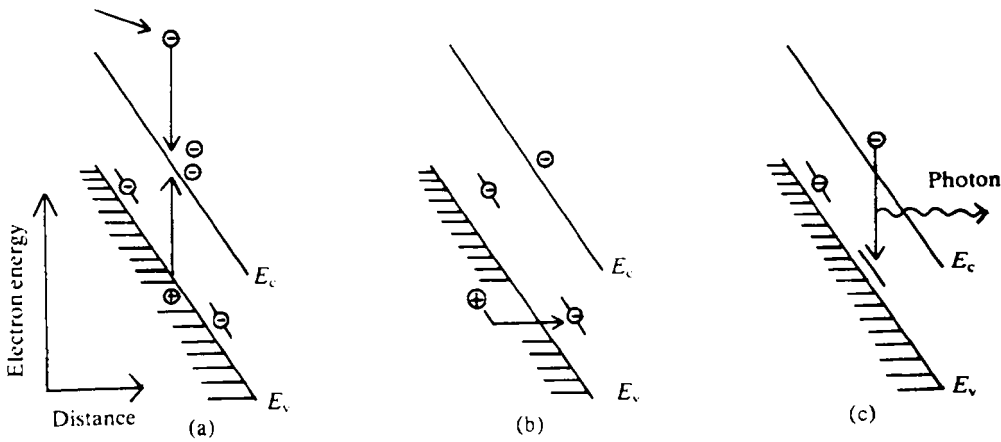


FIG. 4.10 Possible mechanism for electroluminescence emission involving an avalanche process. In (a) an electron moving in the high electric fields present may acquire sufficient energy to excite an electron from the valence band into the conduction band. The hole left behind then moves up into an acceptor state effectively emptying it of an electron (b). Finally, an electron in the conduction band may then make a radiative transfer into the empty acceptor level (c).

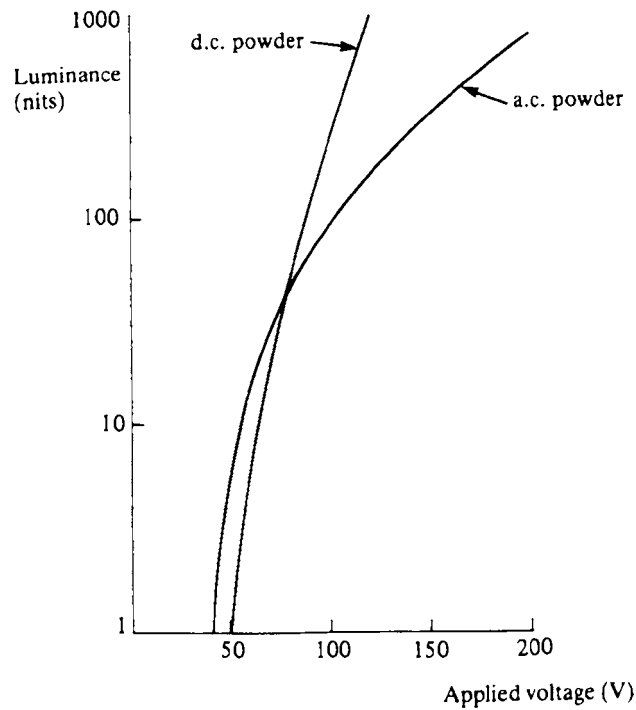


FIG. 4.11 Typical luminances obtained for a.c and d.c. electroluminescent powder devices as a function of the applied voltage (r.m.s. volts in the case of the a.c. device).

basically similar to that of the a.c. device. However, the phosphor particles ($\text{ZnS}:\text{Cu},\text{Mn}$) are coated with a conducting layer of Cu_2S . Provided the phosphor particles are not too widely dispersed within the binder there will be a conducting path from the anode to the cathode. Before normal operation, the cell must be 'formed' by applying a high voltage across it for a short time. This causes copper ions to migrate away from the phosphor surfaces next to the anode. A thin high resistance layer of ZnS is then created next to the anode, across which most of the applied voltage appears, and from which light emission takes place during subsequent operation at lower voltages. Luminances of about 300 nits are possible at voltages of around 100 V d.c., although the power conversion efficiencies are low at approximately 0.1%. The luminance versus drive voltage characteristics for both a.c. and d.c. powder electroluminescent devices are shown in Fig. 4.11.

In addition, a.c. and d.c. devices have been made where the active layer is a vacuum-deposited thin film of phosphor material (usually based on ZnS), but both types tend to have rather poor lifetimes. Although a considerable amount of research effort has been put into the development of electroluminescent displays, they have yet to make any significant commercial impact. They do offer the possibility, however, of large area displays with a high surface area to volume ratio, and their output voltage characteristics, particularly of the d.c. powder devices, are suitable for matrix addressing, as discussed in section 4.10.

4.6

Injection luminescence and the light-emitting diode

The basic structure giving rise to injection luminescence is that of a p-n junction diode operated under forward bias which was discussed in Chapter 2. Under forward bias, majority carriers from both sides of the junction cross the depletion layer and enter the material at the other side, where they are then the minority type of carrier and cause the local minority carrier population to be larger than normal. This situation is described as *minority carrier injection*. The excess minority carriers diffuse away from the junction recombining with majority carriers as they do so. Using eq. (2.42), we may write the excess electron concentration $\Delta n(x)$ in the p material as a function of distance x from the edge of the depletion region as

$$\Delta n(x) = \Delta n(0)\exp(-x/L_c)$$

The process is illustrated in Fig. 4.12. Ideally, in a light-emitting diode (LED) every injected electron takes part in a radiative recombination process and hence gives rise to an emitted photon. In practice this is not so, and the efficiency of the device may be described in terms of the quantum efficiency, which is defined as the rate of emission of photons divided by the rate of supply of electrons. In reverse bias, no carrier injection takes place and consequently no light is emitted. The current-voltage (i , V) relationship for a diode can usually be written (see eq. 2.51 and section 2.8.5) as

$$i = i_0 \left[\exp\left(\frac{eV}{\beta kT}\right) - 1 \right] \quad (4.4)$$

where i_0 is a constant (the reverse saturation current).

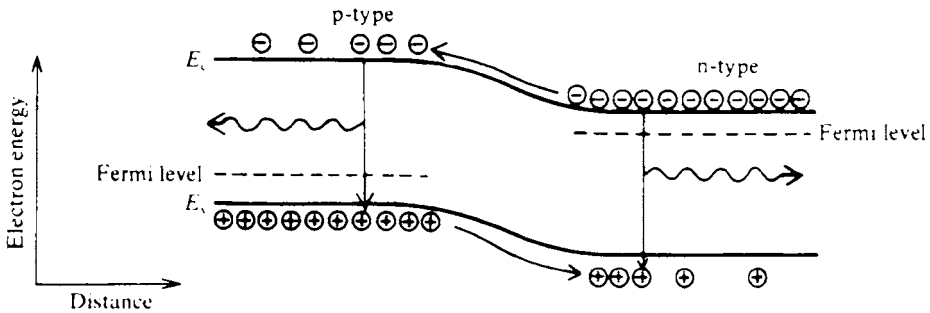


FIG. 4.12 Injection of minority carriers and subsequent radiative recombination with the majority carriers in a forward-biased p-n junction.

The number of radiative recombinations that take place is usually proportional to the carrier injection rate and hence to the total current flowing. If the transitions take place directly between states at the bottom of the conduction band and at the top of the valence band, then the emission wavelength λ_g is given by (see eq. 4.1)

$$hc/\lambda_g = E_c - E_v = E_g$$

Therefore

$$\lambda_g = hc/E_g \quad (4.5)$$

For example, GaAs has an energy bandgap of 1.43 eV, which corresponds to a value for λ_g of 0.86 μm . In fact, because of thermal excitation, electrons in the conduction band have a most probable energy which is $kT/2$ above the bottom of the conduction band (see Problem 2.20). Band-to-band transitions therefore result in a slightly shorter emission wavelength than given by eq. (4.5), and self-absorption can further distort the situation. However, as we shall see later, most transitions involve energy levels within the energy gap, and for these eq. (4.5) represents the shorter wavelength limit. We now consider the transmission process in more detail.

4.6.1 Radiative recombination processes

Radiative recombination in semiconductors occurs predominantly via three different processes, namely (a) interband transitions, (b) recombination via impurity centres and (c) exciton recombination.

4.6.1.1 Interband transitions

This recombination process is illustrated schematically on an energy level diagram for both direct and indirect bandgap materials in Figs 4.13(a) and (b) respectively. ($E-k$ diagrams for silicon and gallium arsenide are shown in Fig. 2.6.) It is important to realize that the transition must conserve the total wavevector of the system. The photon wavevector is given by $2\pi/\lambda$, whilst the electron wavevectors involved range between approximately $-\pi/a$ and

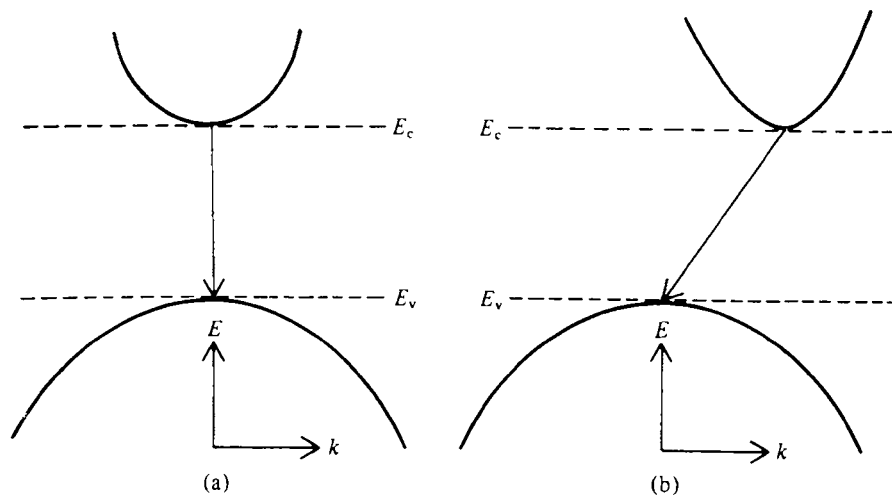


FIG. 4.13 Interband transitions for (a) a direct bandgap semiconductor and (b) for an indirect bandgap semiconductor. In the former case, there is no change in the electron k value, whilst in the latter case there is.

$+\pi/a$, where a is the crystal lattice spacing (i.e. the k values at the first Brillouin zone boundaries; see Fig. 2.5b). For visible radiation, $\lambda \approx 0.5 \times 10^{-6}$ m, whereas crystal lattice spacings are approximately 10^{-10} m; hence $2\pi/\lambda \ll \pi/a$. The photon wavevector is thus much smaller than the magnitude of the maximum possible electron wavevectors. Consequently if the only particles involved are an electron and a photon, then the electron must make a transition between states having virtually the same wavevector. On an $E-k$ diagram, therefore, only vertical transitions are allowed. It is possible to have non-vertical transitions (as illustrated in Fig. 4.13b) but to conserve the wavevector, a phonon must be either created or destroyed at the same time. The equation for the wavelength of the emitted photon is then given by

$$hc/\lambda = E_g \pm E_p \quad (4.6)$$

where E_p is the phonon energy. The $+$ and $-$ signs correspond to phonon *annihilation* and *creation* respectively. Phonon energies are of the order of 0.01 eV, and hence the photon wavelength in fact differs little from λ_g . However, because now *three* particles are involved instead of just two for the direct transition process, the transition is much less probable. We know from the discussion in section 2.4 that the interband recombination rate r may be written

$$r = Bnp \quad (4.7)$$

where B is a constant. Calculated values for B in various semiconductor materials are shown in Table 4.1 (see ref. 4.7), from which we see that the values of B for indirect bandgap materials are some 10^6 times smaller than for direct bandgap materials. We conclude that band-to-band radiative transitions in indirect bandgap semiconductors are relatively rare, and unless some other radiative transition mechanisms are possible these materials will not be suitable for LEDs.

TABLE 4.1 Properties of various semiconductor materials

Group(s)	Element/ compound	Direct/ indirect	E_g (eV)	Readily doped n- or p-type	B ($\text{m}^3 \text{s}^{-1}$)	λ_g (nm)
IV	C	i	5.47			227
	Si	i	1.12	Yes	1.79×10^{-21}	1106
	Ge	i	0.67	Yes	5.25×10^{-20}	1880
IV-VI	SiC (hex. a)	i	3.00	Yes		413
III-V	AlP	i	2.45			506
	AlN	i	5.90	No		210
	AlSb	i	1.50			826
	AlAs	i	2.16			574
	GaN	d	3.40	No		365
	GaP	i	2.26	Yes	5.37×10^{-20}	549
	GaAs	d	1.43	Yes	7.21×10^{-16}	861
	InN	d	2.40			516
	InP	d	1.35	Yes	1.26×10^{-15}	918
	InAs	d	0.35		8.50×10^{-17}	3540
	InSb	d	0.18		4.58×10^{-17}	6870
						387
II-VI	ZnO	d	3.20	No		326
	ZnS(a)	d	3.80	No		344
	ZnS(β)	d	3.60	No		480
	ZnSe	d	2.28	No		544
	ZnTe	d	2.58	No		490
	CdS	d	2.53	No		712
	CdSe	d	1.74	No		826
	CdTe	d	1.50	Yes		

One disadvantage of radiation derived from direct bandgap recombination is that the probability of the emitted radiation being reabsorbed in band-to-band transitions can be high when the radiation has to traverse an appreciable thickness of the semiconductor material.

4.6.1.2 Impurity centre recombination

Three types of recombination involving impurity energy levels are shown in Fig. 4.14. Thus we may have (a) conduction band–acceptor level transitions, (b) donor level–valence band transitions and (c) if in addition a pair of donor and acceptor states are close together, then donor–acceptor transitions are possible. When the electron is in either type of impurity state, then it will be fairly strongly localized. This spatial localization implies that the electron can have a range of momentum values, since uncertainties in both position (Δx) and momentum (Δp) are related via the Heisenberg uncertainty relation (eq. 2.3)

$$\Delta x \Delta p \geq \hbar/2$$

Because the magnitudes of the electron wavevector k and the momentum p are related by

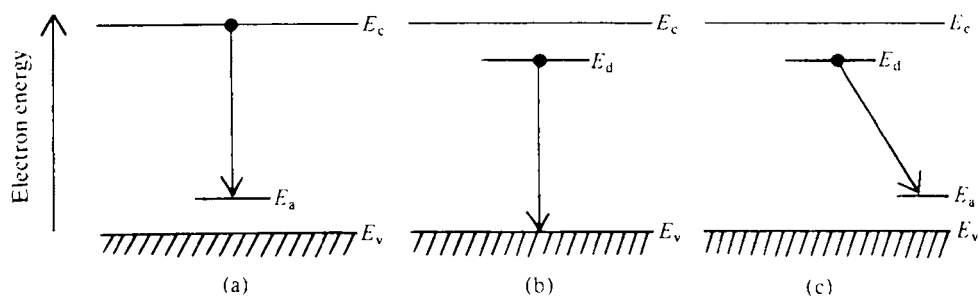


FIG. 4.14 Three types of recombination involving impurity energy levels are shown. In (a) an electron moves from the conduction band into an empty acceptor level. In (b) an electron in a donor level recombines with a hole in the valence band. In (c) an electron in a donor level falls into an empty acceptor level. This latter process requires that the donor and acceptor levels are physically close together.

$k = p/\hbar$ we may write

$$\Delta k \geq \frac{1}{2\Delta x} \quad (4.8)$$

We express the electron localization by putting $\Delta x = Na$, where a is the lattice spacing and N a number which is not expected to be much larger than unity; whence $\Delta k \geq 1/(2Na)$. In our discussion at the start of section 4.6.1.1, we noted that electrons in the conduction and valence bands have a range of k values extending approximately from $-\pi/a$ to $+\pi/a$. Hence we see that there may be sufficient spread in the k values of the impurity states to allow a significant number of transitions between them and the band extrema without calling on the assistance of phonons. Transitions via impurity states therefore provide a possible mechanism whereby indirect bandgap semiconductors can increase their radiative efficiency.

Typical values for $E_c - E_d$ and $E_a - E_v$ are of the order of 0.02 eV, so that the emission wavelength will be slightly longer than that given by eq. (4.5). However, thermal excitation within the bands themselves will tend to decrease the magnitude of this effect (see the discussion following eq. 4.5). Since the radiation has a lower photon energy than that required to excite an electron across the bandgap, it is not subject to reabsorption to the same extent as that derived from band-to-band recombinations.

4.6.1.3 Exciton recombination

Exciton states exist within the energy gaps of even pure semiconductor materials (they must not be confused with impurity donor or acceptor states). We may visualize such states as being akin to Bohr-like states in which an electron and a hole circle round their common centre of gravity at relatively large distances (Fig. 2.12). The electron and hole are relatively weakly bound and the exciton states are situated just below the bottom of the conduction band. In Chapter 2 we showed (eq. 2.28a) that we may write the exciton binding energy E_e as

$$E_e = 13.6 \frac{m_r^*}{m} \left(\frac{1}{\epsilon_r} \right)^2 \text{ eV} \quad (4.9)$$

where m_r^* is the reduced mass. For example, in GaAs we have $\epsilon_r = 11.5$, $m_e^* = 0.068m$ and $m_h^* = 0.47m$, whence $m_r^* = 0.06m$ and $E_c = 5.9$ meV. Observed experimental values are in reasonable agreement with this simple model calculation: in GaAs the exciton binding energy is found to be 4.8 meV. The exciton is capable of movement through the lattice, although, since exciton energies can be affected by the presence of impurities, in some circumstances an exciton may remain 'bound' in the vicinity of the impurity. If the impurities are neutral donors or acceptors then the exciton binding energy is usually about one-tenth of that of the centres to which they are bound. (The binding energy may be much larger than this at isoelectronic traps, which are discussed in section 4.6.3.) Bound exciton states may be sufficiently well localized so that electron-hole recombination can take place in indirect bandgap semiconductors via these states without the need for phonon intervention to conserve the wavevector.

4.6.1.4 Emission linewidths

At a finite temperature electrons in the conduction band of a semiconductor are distributed amongst the energy levels with a probability of occupancy given by the Fermi distribution function. At the end of section 2.5 it is shown that if we denote the number of electrons per unit volume which have energies between E and $E + dE$ by $n(E) dE$ then

$$n(E) = \frac{4\pi}{h^3} (2m_e^*)^{3/2} (E - E_c)^{1/2} \exp\left(-\frac{E - E_F}{kT}\right)$$

The function $n(E)$ is sketched in Fig. 2.14(c), and it can be shown (Problem 2.20) that it has a halfwidth of about $2kT$ (the actual value is close to $1.8kT$). A similar result can be obtained for holes in the valence band. Thus assuming direct band-to-band transitions we would expect a range of emission frequencies, $\Delta\nu$, which have a halfwidth of about $2kT/h$. This may be converted to the equivalent wavelength spread $\Delta\lambda_0$ by use of the relationship $\lambda_0\nu = c$, so that

$$\frac{d\nu}{d\lambda_0} = -\frac{c}{\lambda_0^2}$$

and hence

$$\Delta\lambda_0 = \frac{2kT\lambda_0^2}{hc} \quad (4.10)$$

At a wavelength of 900 nm and a temperature of 300 K, eq. (4.10) predicts a halfwidth of 38 nm which agrees reasonably well with what is observed in practice (see e.g. Fig. 4.15). The result remains reasonably valid even when the transitions involve impurity states rather than direct band-to-band transitions.

4.6.2 LED materials

We may summarize the main requirements for a suitable LED material as follows: first, it must have an energy gap of appropriate width; secondly, both p- and n-types must exist, preferably with low resistivities; and finally, efficient radiative pathways must be present.

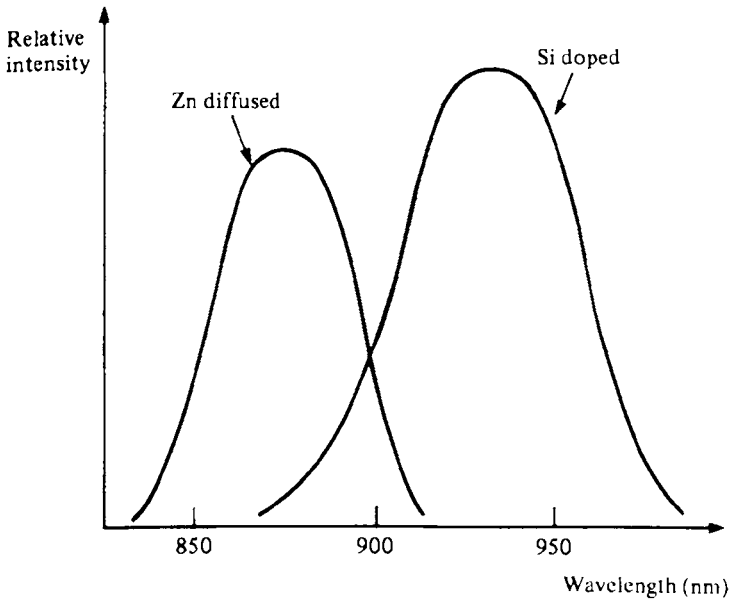


FIG. 4.15 Spectral output of an Si-doped and a Zn-diffused GaAs LED.

Equation (4.5) indicates that, in order to obtain visible radiation, energy gaps greater than or equal to about 2 eV are required. Unfortunately, materials with such large gaps tend to have high resistivities even when doped. Furthermore, in most cases the wider the energy gap, the greater are the difficulties met in material preparation. This is often because the materials have high melting temperatures and low structural stability.

4.6.3 Commercial LED materials

GALLIUM ARSENIDE (GaAs)

This is a direct bandgap semiconductor with $E_g = 1.443$ eV ($\lambda_g = 860$ nm); suitable p-n junctions may be made by diffusing zinc into crystals of n-type GaAs. The resulting radiation arises from band-to-band transitions, however, and is consequently subject to heavy reabsorption; this reduces the device efficiency and shifts the peak emission wavelength to about 870 nm. More efficient diodes may be made by using silicon as a dopant, where, depending on the growth conditions, either p or n material is obtained. Complex acceptor levels are also formed about 0.1 eV above the valance band. Transitions between these and the conduction band give rise to radiation with a peak emission wavelength of about 1000 nm which is not subject to reabsorption. Typical emission spectra for both Zn-diffused and Si-doped diodes are shown in Fig. 4.15.

GALLIUM PHOSPHIDE (GaP)

This is an indirect bandgap semiconductor with $E_g = 2.26$ eV ($\lambda_g = 549$ nm) and hence band-

to-band transitions are rare. Group V elements such as N and Bi may be used as dopants to assist radiative transitions. These replace the phosphorus atoms and form recombination centres called *isoelectronic traps*. In the case of nitrogen the effective trap depth below the conduction band is small (8 meV) and the subsequent radiation has a peak wavelength only slightly less than λ_g . Using increased levels of nitrogen doping and also doping with zinc and oxygen simultaneously give rise to deeper traps and consequently higher emission wavelengths.

GALLIUM ARSENIDE PHOSPHIDE ($\text{GaAs}_{1-x}\text{P}_x$)

The energy gap of this ternary alloy depends on the value of x and furthermore it changes from being a direct bandgap when $x < 0.45$ to an indirect bandgap when $x > 0.45$. Using diodes with $x = 0.4$ results in red emission (Fig. 4.16). The indirect bandgap material can also be used in conjunction with the same radiative assisting dopants as used in GaP.

GALLIUM ALUMINIUM ARSENIDE ($\text{Ga}_x\text{Al}_{1-x}\text{As}$)

Highly efficient red- and near-infrared-emitting LEDs can be made from this material. If a heterojunction is formed between n-type $\text{Ga}_{0.3}\text{Al}_{0.7}\text{As}$ and p-type $\text{Ga}_{0.6}\text{Al}_{0.4}\text{As}$, electrons injected from the n surface layer into the p material recombine radiatively via acceptor levels and result in radiation of 650 nm wavelength. This can pass through the surface layer with little attenuation because of the relative large bandgap of the latter.

III-V NITRIDES (e.g. GaN and AlN)

These materials have energy gaps that correspond to emission wavelengths from green all the way into the ultraviolet. A number of problems frustrated the early development of these materials; for example, until recently there was a lack of suitable substrate materials with matching lattice constants and thermal expansion coefficients. It also proved difficult to dope the materials p-type. It was not until the early 1980s that these problems were overcome, when it was found that the deposition of buffer layers allowed growth on readily available substrates such as sapphire and silicon carbide. In addition new growth and irradiation techniques have enabled p-type layers to be fabricated.

INDIUM GALLIUM ARSENIDE

High brightness LEDs are now commercially available based on InGaAs, although the structures (e.g. 'double heterostructures', see sections 2.8.6 and 5.10.2.3) are somewhat more complicated than those discussed so far. By increasing the indium content the emission wavelength increases, and radiation in the green can also be obtained with reasonably high efficiency.

II-VI SEMICONDUCTORS (e.g. zinc selenide and related compounds)

These materials have also been used successfully to make blue- and green-emitting diodes. For example, green-emitting diodes have been made using ZnTeSe as an active region grown on a ZnSe substrate. By replacing the ZnTeSe layer with one of ZnCdSe emission has been obtained in the blue. However, these materials are much softer than the III-V nitrides and degrade more rapidly at elevated temperatures and consequently have shorter working lifetimes.

TABLE 4.2 Characteristics of the most commonly used LED materials

Material	Dopant	Peak emission (nm)	Colour	External quantum efficiencies (%)
GaAs	Si	910 → 1020	Infrared	10
Ga _x Al _{1-x} As (1 < x < 0.7)	Si	879 → 890	Infrared	15
GaP	Zn, O	700	Red	4
GaAs _{0.6} P _{0.4}		650	Red	0.2
Ga _{0.6} Al _{0.4} As	Zn	650	Red	15†
GaAs _{0.45} P _{0.55}	N	632	Orange	0.2
GaP	N, N	590	Yellow	0.1
AlInGaP		570	Yellow	1†
GaP	N	555	Green	0.1
Zn _{0.9} Cd _{0.1} Se		489	Blue	1.3†
SiC		470	Blue	0.03
In _{0.06} Ga _{0.94} As	Zn	450	Blue	3.8†

† Indicates a double heterostructure diode.

SILICON CARBIDE (SiC)

Silicon carbide has been a promising material for many years, but its high melting point and consequent growth difficulties prevented its early use in LEDs. It can be doped as both n- and p-type and commercial blue-emitting diodes have been available for some years. Doping with B, Al, Sc and Be gives rise to yellow, blue, green and red emission respectively. However, efficiencies have always been very low and with the recent development of highly efficient InGaAs emitters the future prospects for this material do not seem too good.

Table 4.2 summarizes the characteristics of the most commonly used LED materials.

4.6.4 LED construction

A typical LED construction is shown in Fig. 4.16. It is obviously advantageous if most of the radiative recombinations take place from the side of the junction nearest the surface, since then the chances of reabsorption are lessened. We may ensure this by arranging that most of the current flowing across the diode is carried by those carriers that are injected into the surface layer. We assume that, as in Fig. 4.16, the surface layer is p-type. The fraction of the total diode current that is carried by electrons being injected into the p side of the junction (η_e) is then given by

$$\eta_e = \frac{D_e n_p / L_e}{D_e n_p / L_e + D_h p_n / L_h} \quad \text{or} \quad \eta_e = \left(1 + \frac{D_h L_e p_n}{D_e L_h n_p} \right)^{-1} \quad (4.11)$$

This equation is readily derived from the results of section 2.8.2. By using the Einstein relation (eq. 2.40), $D_{e,h} = (kT/e)\mu_{e,h}$, and the relation $n_p p_p = p_n n_n = n_i^2$ (eq. 2.36), eq. (4.11)

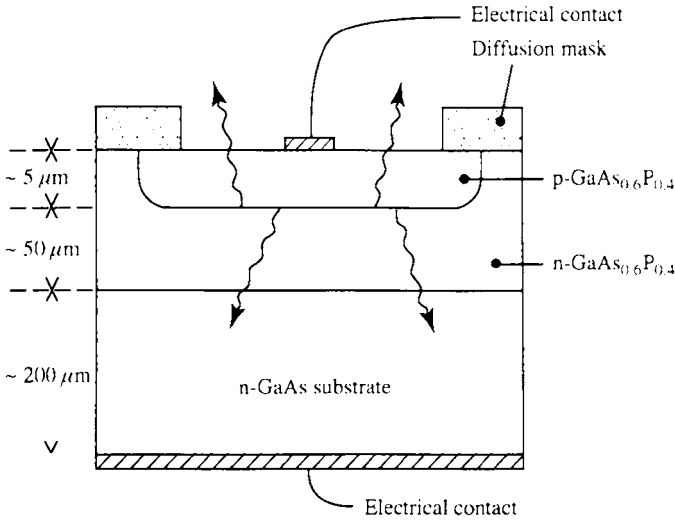


FIG. 4.16 Cross-section of a standard red-emitting LED chip based on GaAsP. A layer of n-type GaAs_{0.6}P_{0.4} (using tellurium as a dopant) is deposited by vapour phase epitaxy on a GaAs substrate. A p-n junction is then formed by diffusing in Zn through a surface mask. The small aluminium contact on the upper surface allows as much of the radiation to escape as possible; any radiation flowing downwards is almost completely absorbed by the GaAs.

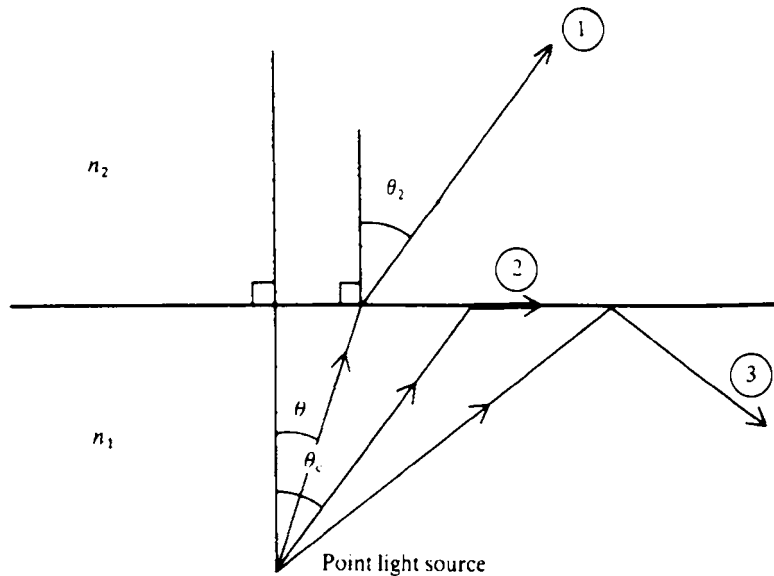


FIG. 4.17 Phenomenon of total internal reflection. When a beam of light is incident at an angle θ_1 onto an interface between two media of refractive indices n_1 and n_2 ($n_2 < n_1$) then the refracted beam (beam 1) makes an angle θ_2 with the normal to the surface where $n_1 \sin \theta_1 = n_2 \sin \theta_2$ (Snell's law). At the critical angle ($\theta_1 = \theta_c$), $\theta_2 = 90^\circ$ (beam 2) and the refracted beam emerges along the interface. At angles of incidence greater than θ_c (beam 3), the beam is totally reflected back into the first medium.

becomes

$$\eta_e = \left(1 + \frac{\mu_h p_p L_c}{\mu_e n_n L_h} \right)^{-1} \quad (4.12)$$

In III–V compounds $\mu_e \gg \mu_h$, and so, assuming that $L_c \approx L_h$, we see that there is a natural tendency for η_e to be close to unity. We may reinforce this tendency by arranging that $n_n \gg p_p$ (i.e. by making an $n^+ - p$ diode).

Although the internal quantum efficiencies of some LED materials can approach 100%, the external efficiencies are much lower. The main reason for this is that most of the emitted radiation strikes the material interface at an angle greater than the critical angle and so remains trapped. Unfortunately, the high refractive indices of the III–V materials discussed here give rise to small critical angles. Consider, for example, radiation from a point source within a medium of refractive index n_1 impinging on a plane interface with another medium of refractive index n_2 , where $n_2 < n_1$, as shown in Fig. 4.17. Only those rays (e.g. beam 1) that have an angle to the normal less than the critical angle (θ_c) enter the second medium. Those with angles greater than θ_c (e.g. beam 3) are reflected back into the first medium.

From eq. (1.14), we have that the critical angle θ_c is given by

$$\theta_c = \sin^{-1}(n_2/n_1) \quad (4.13)$$

Light originating at recombination centres near the $p - n$ junction will be radiated isotropically, whereas only that within a cone of semiangle θ_c will escape. In Problem 4.1, it is shown that the fraction F of the total generated radiation that is actually transmitted into the second

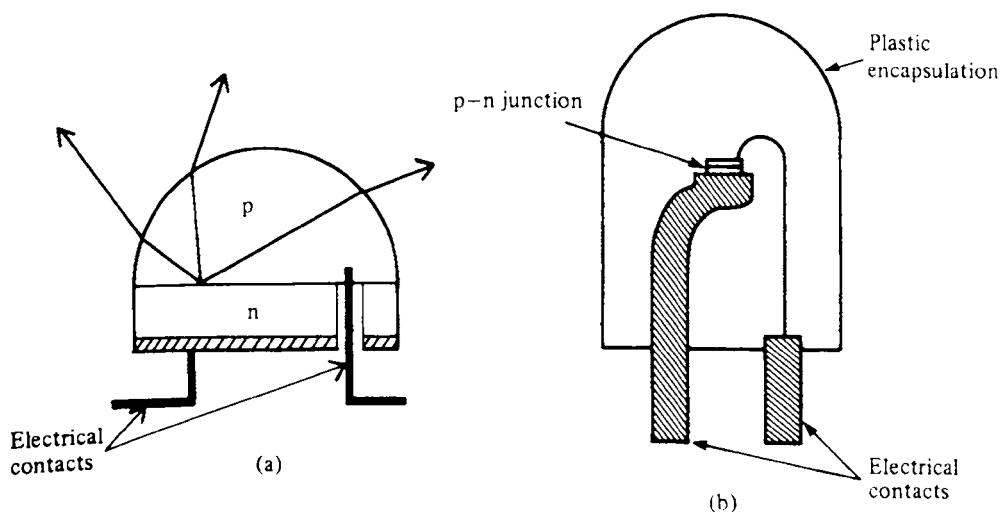


FIG. 4.18 Two methods used to reduce reflection losses in LEDs. In (a) the p material is made into a hemispherical dome. More radiation then strikes the semiconductor/air interface at less than the critical angle than would otherwise be the case. In (b) the $p - n$ junction is surrounded by plastic encapsulation. Losses at the plane semiconductor/plastic interface are then less than for a corresponding semiconductor/air interface.

medium is

$$F \approx \frac{1}{4} \left(\frac{n_2}{n_1} \right)^2 \left[1 - \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \right] \quad (4.14)$$

There are two obvious ways to increase F ; the first is to ensure that most rays strike the surface at less than the critical angle. This may be achieved by shaping the semiconductor/air interface into a hemisphere, as shown in Fig. 4.18(a). However, although this technique is used occasionally in high power diodes, it is too difficult and expensive for most situations. The second, and much commoner, technique is to encapsulate the junction in a transparent medium of high refractive index. This is usually a plastic material with a refractive index of about 1.5. Using eq. (4.14) with $n_1 = 3.6$ and $n_2 = 1.5$ we obtain $F = 0.036$, giving a nearly threefold increase in light output over the simple semiconductor/air interface (Example 4.2). Of course there will be some losses at the plastic/air interface, but these are easily minimized by moulding the plastic into an approximately hemispherical shape (see Fig. 4.18b).

The diodes themselves may be fabricated using either vapour or liquid phase epitaxy; for more details the reader is referred to ref. 4.8.

EXAMPLE 4.2 Transmission efficiency of a plane GaAs LED surface

If we take a GaAs/air interface where $n_1 = 3.6$ and $n_2 = 1$, then the fractional transmission for isotropic radiation originating inside the GaAs is given by eq. (4.14) as

$$\begin{aligned} F &\approx \frac{1}{4} \left(\frac{1}{3.6} \right)^2 \left[1 - \left(\frac{2.6}{4.6} \right)^2 \right] \\ &\approx 0.013 \end{aligned}$$

The basic reason for this low transmission efficiency for a GaAs/air interface may be better appreciated if we calculate the critical angle. From eq. (4.13) we have $\theta_c = \sin^{-1}(1/3.6) = 16^\circ$.

4.6.5 Response times of LEDs

For most display purposes the fast response time obtainable from LEDs ($< 1 \mu\text{s}$) is not essential. However, because of their importance in optical communications (see Chapter 9), it is convenient to pursue the matter here. There are two main factors which limit the speed with which an LED can respond to changes in the drive current. One of these is due to the effects of junction capacitance C_j which arises from the variation in charge stored in the depletion region when the external voltage is varied. It was shown in section 2.8.4 that C_j varies as the reciprocal of the square root of the applied voltage (assuming an 'abrupt' junction).

The other limitation is due to what is sometimes termed 'diffusion capacitance', and results from the storage of mobile carriers within a diffusion length or so of the junction. When the external voltage is reduced, charge must diffuse away from the junction and disappear by recombination to enable the new equilibrium conditions to be attained. It is shown in

Appendix 3 that the frequency response resulting from this process may be written

$$R(f) = \frac{R(0)}{(1 + 4\pi^2 f^2 \tau_c^2)^{1/2}} \quad (4.15)$$

Here $R(f)$ is the response at frequency f and τ_c is the minority carrier lifetime provided we have conditions of low level injection. For high level injection the concept of a constant lifetime no longer applies (see the arguments leading to eq. 2.38), and we must assume some average value of τ_c . However, in practice eq. (4.15) is found to be a good representation of the frequency response of most LEDs. It is evident that for a good frequency response we require τ_c to be as small as possible. From eq. (2.38) we have for p-type material that

$$\tau_c = (Bp)^{-1} \quad (4.16)$$

where p is the majority carrier population (here assumed to be holes) and B is the constant tabulated in Table 4.1. We see that τ_c may be reduced by using highly doped material. Unfortunately if compounds such as GaAs are doped at near the solubility limit for acceptor impurities, then non-radiative centres are formed. Germanium is a fairly commonly used acceptor impurity in GaAs and above a concentration of some 10^{24} atoms m^{-3} the external quantum efficiency starts to decline. At this concentration, the electron lifetime is given by eq. (4.16):

$$\tau_c = (7 \times 10^{-16} \times 10^{24})^{-1} = 1.4 \times 10^{-9} \text{ s}$$

An alternative approach to attaining short response times is to use lightly doped material with a narrow active region and to operate the diode under conditions of heavy forward injection. The injected electron and hole densities are then much greater than they are in equilibrium. If Δn and Δp are the injected carrier concentrations of electrons and holes, then eq. (4.16) can be written

$$\tau_c = (B\Delta p)^{-1} \quad (4.17)$$

If the injection current density is J and the active region width is t , then in equilibrium the number of recombinations per second per unit volume must be J/te . Since an excess population density Δp gives rise to $\Delta p/\tau_c$ recombinations per second per unit volume, we have

$$\Delta p = \frac{J\tau_c}{et} \quad (4.18)$$

Eliminating Δp from eqs (4.17) and (4.18), we obtain

$$\tau_c = \left(\frac{et}{JB} \right)^{1/2} \quad (4.19)$$

We see that in this case τ_c may be reduced by reducing t and increasing J . However, since the lifetime is now current dependent, this approach can lead to signal distortion.

4.6.6 LED drive circuitry

As we have seen, the electrical characteristics of LEDs are essentially those of ordinary rectifying diodes. Typical operating currents are between 20 mA and 100 mA, whilst the forward

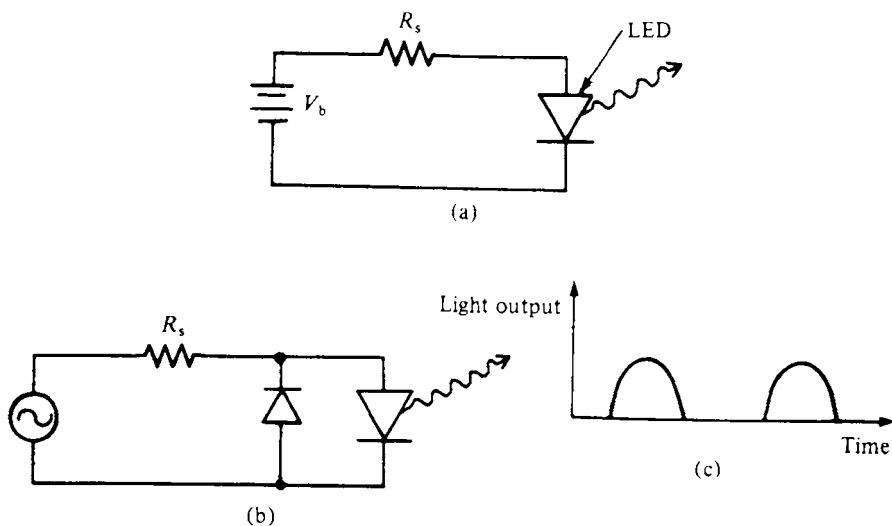


FIG. 4.19 Simple LED drive circuits for (a) d.c. operation and (b) a.c. operation. In both cases a series resistor R_s limits the maximum current flow. In the a.c. circuit a diode is placed with reversed polarity across the LED to prevent damage from excessive reverse bias voltages; (c) shows the light output obtained with the circuit of (b).

voltages vary from 1.2 V for GaAs to 2 V for GaP. (The operating voltage is approximately equal to the built-in diode potential, which in turn is slightly less than the energy gap expressed in eV.) Simple drive circuits for d.c. and a.c. voltage operation are shown in Figs 4.19(a) and (b) respectively. The current through the diode is limited by a series resistance R_s whose value may be calculated from

$$R_s = \frac{V_b - V_d}{i_d} \quad (4.20)$$

where V_b is the power source voltage, V_d the diode operating voltage and i_d the desired diode current. In the a.c. circuit, a rectifying diode is placed across the LED to protect it against reverse bias breakdown.

These two circuits provide for continuous 'on' operation. If it is desired to switch the diode on or off, or to modulate the output, then the circuits shown in Figs 4.20(a) and (b) respectively may be used. In Fig. 4.20(a) the transistor is used as a simple switch. With no voltage applied to the base, the transistor has a very high impedance between the collector and emitter and hence no current flows through the LED. If a large enough base voltage is then applied so that the emitter-base junction becomes heavily forward biased, the transistor has a relatively low impedance between emitter and collector and a substantial current can flow, resulting in the LED being turned on. In Fig. 4.20(b) the transistor is biased so that the quiescent diode current is about half its peak value and both the transistor and the LED are biased well into their linear regions. Changes in the current flowing through the LED are then directly proportional to changes in the input voltage.

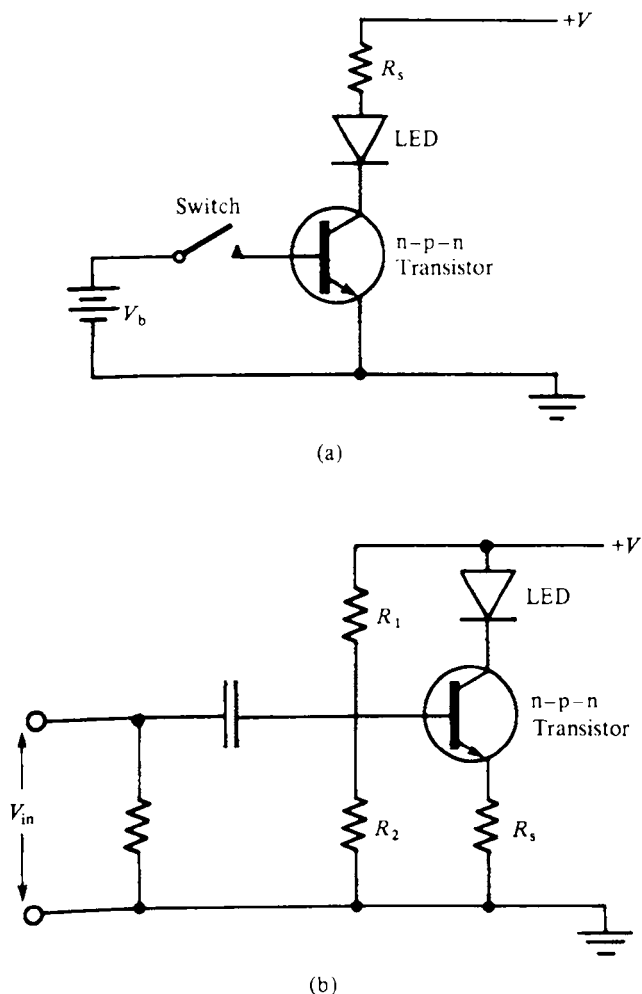


FIG. 4.20 LED modulation circuits: (a) provides for a simple on/off modulation via a switch. The voltage V_b is sufficient to switch the transistor on. There is then a low impedance path between the collector and emitter terminals, and the current flowing through the LED is determined by the voltage V and the series resistance R_s . In (b) the diode output may be modulated by voltage V_{in} . The resistors R_1 and R_2 bias the transistor so that the average current through the transistor, and hence through the LED, is about half the maximum value. Both the transistor and the LED are then biased well into their linear regions.

4.7

Plasma displays

Plasma displays rely on the glow produced when an electrical discharge takes place in a gas (usually neon). Free electrons present in the discharge acquire high kinetic energies from the external field and when they collide with the gas atoms (or ions) they transfer this energy

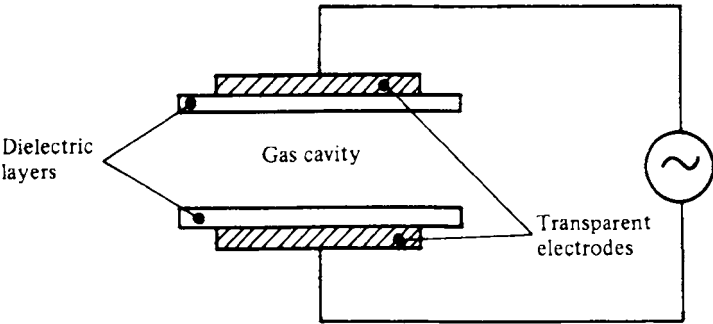


FIG. 4.21 Construction of an a.c. plasma element. The gas cavity is some 10^{-4} m in width with transparent electrodes on the outside of the containing dielectric layers. The main constituent of the gas is

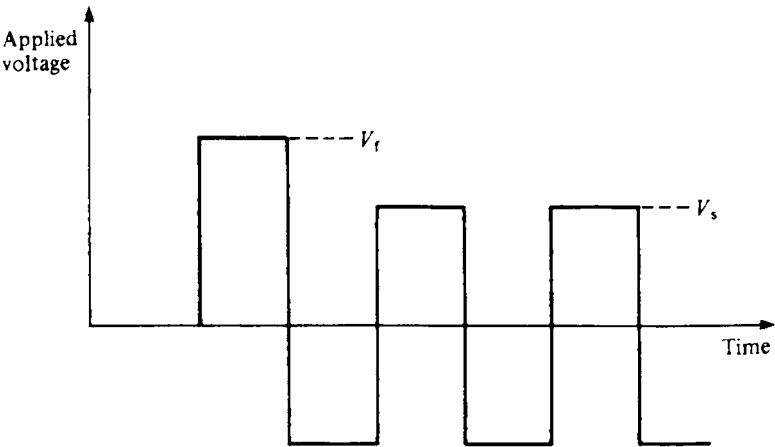


FIG. 4.22 Voltage waveform used for driving an a.c. plasma display. After an initial firing pulse V_f the voltage may be reduced to V_s for the remaining time the display is required to be on.

to the atoms, thereby exciting them into energy levels above the ground state. The atom may then lose energy radiatively and return to the ground state.

Either a.c. or d.c. excitation may be used, although the former is the most common. Figure 4.21 shows the basic construction of an a.c. plasma display element in which the electrodes are external to the gas cavity. (When d.c. excitation is used, the electrodes must be inside the gas cavity.) Typical cavity widths and gas pressures are 100 μm and 400 torr respectively. The discharge is initiated by applying a firing voltage V_f of some 150 V; however, once the discharge has started, it may be sustained with a reduced voltage V_s of about 90 V. A suitable voltage waveform for this situation is shown in Fig. 4.22.

If the initial voltage pulse is relatively wide then an appreciable amount of charge accumulates on the interior of the dielectric layer. This charge build-up causes an internal electric field to be set up which, being in opposition to the external field, may be strong enough

to extinguish the discharge if the pulse width is too long. Such accumulated charge takes a long time to dissipate and this effect can be used to enable the device to have a memory. Thus, if at some time subsequent to the first ('writing') pulse an external voltage pulse of opposite polarity is applied, the field from the stored charge will add to the external field and the discharge will again commence, but now at a lower voltage than V_p .

4.8 Display brightness

The basic photometric units of brightness were introduced in section 1.4, and we briefly review the topic again here. The sensitivity of the human eye to different wavelengths was shown in Fig. 1.24, which enables us to convert objective radiometric units into subjective photometric units. Thus, if the radiant power in the wavelength interval λ to $\lambda + d\lambda$ (nm) is $W(\lambda) d\lambda$ (watts), then the total luminous flux Φ (lumens) is given by

$$\Phi = 680 \int_{380}^{780} W(\lambda) V_\lambda d\lambda \quad (4.21)$$

where V_λ is the relative response of the eye as shown in Fig. 1.17. This function has its maximum at a wavelength of 555 nm, where the eye's absolute sensitivity is 680 lm W^{-1} .

A calculation of Φ enables the conversion efficiency (in lm W^{-1}) of electrical power into effective visible radiation to be determined. However, the apparent brightness of a display element depends on the emission area and the angle from which it is observed. If the element is sufficiently small for it to appear as a point source (which is the case if it subtends an angle at the eye of some 2 minutes of arc or less), then a meaningful measure of brightness is the flux per unit solid angle in the direction of the observer. This is called the *luminous intensity* of a source and the units are *lumens per steradian* or *candela*. For sources with an apparently finite area, the brightness is more properly referred to as the *luminance* and the appropriate units are *lumens per steradian per unit projected area of emitting surface*, that is *candelas m⁻²* or *nits*.

The variation in luminance with viewing angle is something that must be determined experimentally for each type of display. It is often assumed that emitters show either isotropic emission or Lambertian emission. If the luminance of a surface viewed at an angle θ to the normal is $B(\theta)$, then for isotropic emission $B(\theta) = \text{constant} = B(0)$. For Lambertian emission, on the other hand, $B(\theta) = B(0) \cos(\theta)$. It is readily shown (see Problem 4.5) that for an isotropic surface of area A the total flux Φ_i is given by

$$\Phi_i = 2\pi AB(0) \quad (4.22)$$

whilst for a corresponding Lambertian surface

$$\Phi_L = \pi AB(0) \quad (4.22a)$$

Quite a number of displays are Lambertian, but there are some, for example GaP red LEDs, that approximate more to isotropic emitters.

EXAMPLE 4.3 Brightness of an LED

We take an LED with a chip diameter of 0.2 mm which is viewed from a distance of 1 m. It emits at a wavelength of 550 nm and has an external quantum efficiency of 0.1%. We further assume that the emission is isotropic and that the diode is operated at 50 mA. We must first decide whether the diode acts as a point or as an extended source at the eye. The emitting area subtends an angle θ where

$$\tan \frac{\theta}{2} = \frac{2 \times 10^{-4}}{2 \times 1}$$

whence $\theta < 1$ minute of arc, and the LED acts as a point source.

The total radiant power output W is given by

$W = hc/\lambda \times \text{quantum efficiency} \times \text{photon emission rate}$

$$= \frac{hc}{550 \times 10^{-9}} \times (0.001) \times \frac{(50 \times 10^{-3})}{e}$$

Therefore $W = 1.13 \times 10^{-4}$ watts.

Using Fig. 1.24, we take an average luminosity at 550 nm of 600 lm W^{-1} ; hence the luminous flux from the source is $1.13 \times 10^{-4} \times 600$, that is $6.8 \times 10^{-2} \text{ lm}$. Isotropic emission means the flux is uniformly distributed over the solid angle 2π , and so the luminous intensity at normal incidence is $6.8 \times 10^{-2}/2\pi = 1.1 \times 10^{-2}$ candelas.

Another brightness unit sometimes encountered is the foot-Lambert (ft-L). This is obtained by dividing the total luminous flux (in lumens) by the area of the device in square feet. This definition assumes a Lambertian emission pattern. To convert nits or candelas m^{-2} into foot-Lamberts (for a Lambertian surface) we multiply by the factor 0.292 (this factor may be derived from eq. 4.22a).

The brightness level required of a display will depend on the ambient lighting conditions. In a dimly lit room, a brightness of about 10 nits may suffice, whilst in a well-lit environment about 1000 nits may be needed, although the difference between the display on and off brightnesses also plays an important role. A more thorough discussion is given in ref. 4.9.

4.9**Liquid crystal displays**

We turn now to the most important of the 'passive' types of display (and indeed the only one of this type we shall mention), namely liquid crystal displays (LCDs). These have come into prominence in the last few years mainly as display elements for digital watches and pocket calculators. Here one of the prime requisites is for low power consumption, particularly so for the digital watch because of the necessarily low capacity of the power source. LCDs consume the least power of all common display devices because no light generation is required. This, as we have seen in the LED, for example (Table 4.2), is a very inefficient process; at best the efficiency is only some 10%. There are two basic types of LCD avail-

able. These are (a) reflective, which requires front illumination, and (b) transmissive, which requires rear illumination. Most reflective types utilize ambient light for illumination with provision for secondary illumination via a small incandescent lamp or LED if ambient levels become too low. At the heart of all LCD devices is a cell formed between two glass plates, each with a conductive coating. The cell has a thickness of about 10 μm (sometimes less) and is filled with a liquid crystal material.

The liquid crystal state is a phase of matter which is exhibited by a large number of organic materials over a restricted temperature range. At the lower end of the temperature range, the material becomes a crystalline solid, whilst at the upper end it changes into a clear liquid. Within this range it has a milky yellowish appearance and combines some of the optical properties of solids with the fluidity of liquids. A major characteristic of all liquid crystal compounds is the rod-like shape of their molecules. When they are in the liquid crystal phase, these molecules can take up certain orientations relative both to each other and to the liquid crystal surface. It is usual to describe this orientation in terms of a *director*, that is a unit vector pointing along the time-averaged preferred orientation of the molecules in any small volume.

There are three basic types of ordering in liquid crystals, which are termed *nematic*, *cholesteric* and *smectic*. Only the first two of these are of importance in display devices at present and are illustrated in Fig. 4.23. In nematic ordering, the molecules (or, rather, the directors) are aligned parallel to each other, but apart from remaining parallel the molecules are free to move relative to each other so that the phase has liquid properties. A nematic liquid crystal molecule usually consists of two benzene rings linked with a central group. A typical example is 4-methoxybenzylidene-4-butylaniline (MBBA), which has the chemical formula $\text{CH}_3\text{—O—}\text{C}_6\text{H}_4\text{—CH=N—C}_6\text{H}_4\text{—C}_4\text{H}_9$. MBBA shows liquid crystal behaviour over the temperature range 20°C to 47°C. For more general details concerning liquid crystal materials the reader is referred to ref. 4.10.

In the cholesteric phase (Fig. 4.23b), we may regard the material as being made up from a large number of planes each having a nematic-like structure, but with each plane showing a progressive change in the director direction from the one below. The director directions thus display a helical twist through the material. The distance between planes having the same director direction is called the *pitch*, p . Cholesteric liquid crystals exhibit some interesting colour effects. If, for example, light of wavelength λ is incident normally on the director planes, then strong Bragg reflection will occur when $p = m\lambda$ (m an integer), but not otherwise. Thus if white light is shone onto a cholesteric liquid crystal, it can appear strongly coloured. Furthermore, the pitch is usually temperature dependent, so that the colour of the reflected light will also be temperature dependent. Obviously this can form the basis of a thermometer. Most liquid displays, however, are of the so-called twisted nematic type, and we shall spend the rest of this section discussing them.

When a nematic liquid crystal material comes into contact with a solid surface, the directors often become aligned either perpendicular to the surface (*homeotropic* ordering) or parallel to the surface (*homogeneous* ordering). These two forms can be produced by suitable treatment of the surface. In the case of homogeneous ordering, this can often be achieved by rubbing the surface once or twice along a particular direction with a soft fabric (e.g. cotton) before it comes into contact with the liquid crystal material. The liquid crystal directors then take up an orientation parallel to the direction of rubbing.

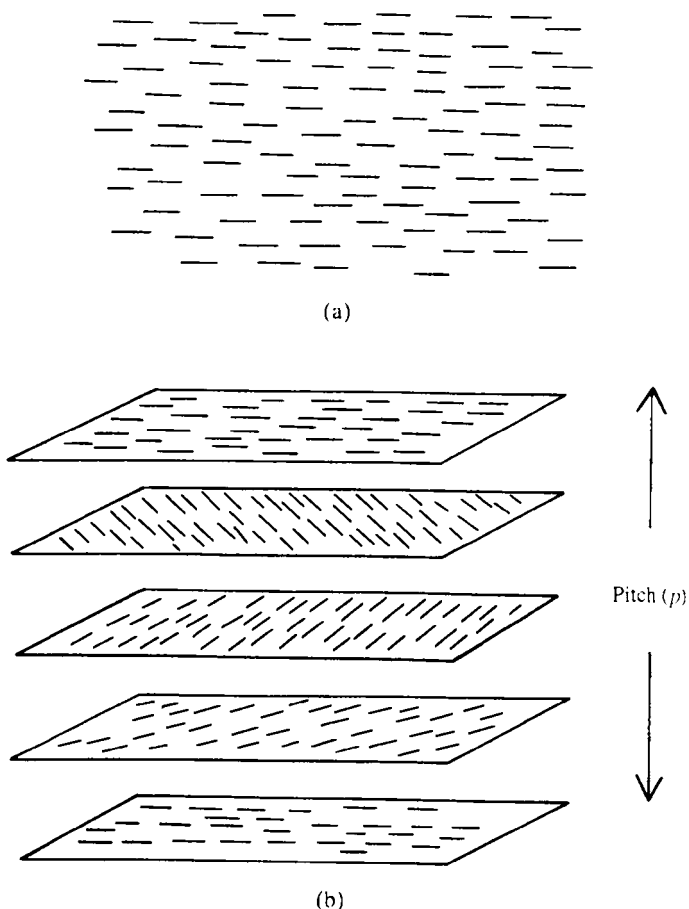


FIG. 4.23 (a) Nematic ordering and (b) cholesteric ordering. In (a) the directors all line up parallel to each other. In (b) a large number of planes of nematic order are formed where the directors rotate as we move along a direction perpendicular to the planes.

One of the most important electrical characteristics of liquid crystal materials is that they show different dielectric constants ϵ_{\parallel} and ϵ_{\perp} , depending on whether the external field is parallel to, or perpendicular to, the molecular axis. If $\epsilon_{\parallel} > \epsilon_{\perp}$ we refer to it as a *positive* material. The application of an external electric field to a positive material will tend to make the molecules lie along the electric field, since this will tend to minimize their energy. We see that there is thus a possibility of changing the homogeneous type of ordering into a homeotropic type by the application of a field which is perpendicular to the surface (assuming a positive material). This transition is found to take place above a critical field (\mathcal{E}_c) and is illustrated in Fig. 4.24.

The most common liquid crystal display uses a 'twisted nematic' cell. In this, the opposite walls of the cell are treated to produce a homogeneous arrangement in which the molecular alignment directions at the walls are at right angles to each other. Thus the molecules

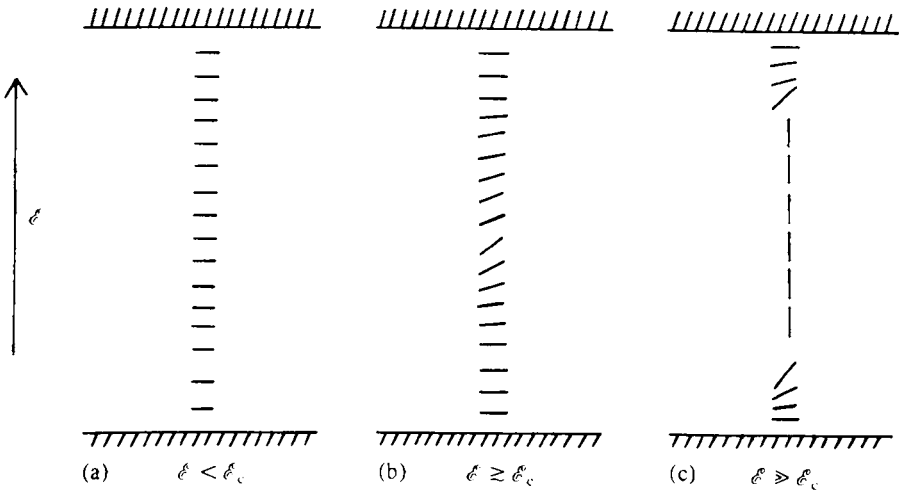


FIG. 4.24 Behaviour of molecules in an initially homogeneously ordered liquid crystal material as an increasing electric field of magnitude E is applied in a direction perpendicular to the liquid crystal/solid interface. If E is less than a critical value (E_c) the ordering is not affected (a). If $E \geq E_c$, the molecules furthest away from the interface begin to align along the field direction (b). If $E \gg E_c$ then most of the molecules are aligned along the field direction (c).

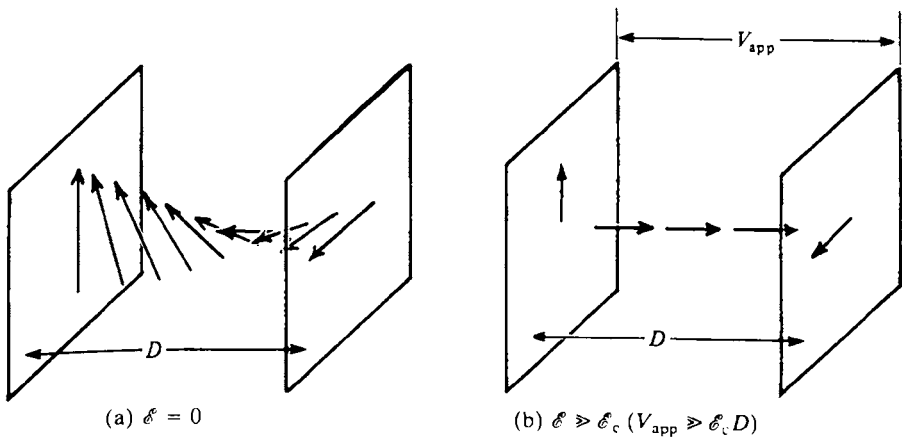


FIG. 4.25 Behaviour of the molecules in a liquid crystal cell of thickness D with (a) no applied voltage ($E = 0$) and (b) with a voltage applied such that $E \gg E_c$. In the former the molecules undergo a 90° rotation across the cell, in the latter they are mostly ordered with their axes parallel to the applied field.

undergo a 90° rotation across the cell as shown in Fig. 4.25(a). When a beam of polarized light is incident on the cell the strong optical anisotropy of the liquid causes the polarization to undergo a 90° rotation. With a strong enough electric field across the cell, however (i.e. $E \gg E_c$), the molecular alignments will become as shown in Fig. 4.25(b) and in this state the molecular alignments will have no effect on an incident polarized light beam.

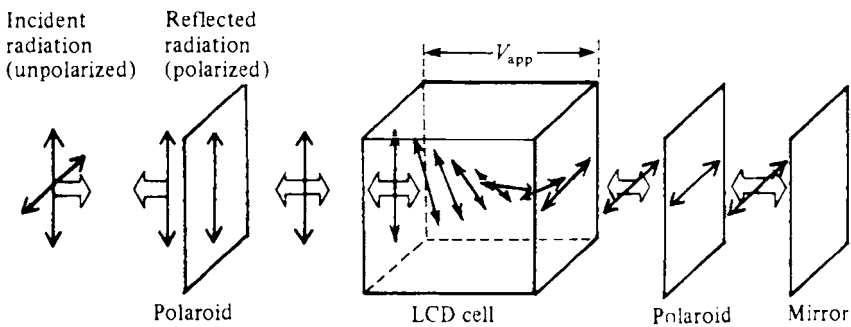


FIG. 4.26 Illustration of the action of an LCD display device.

In operation, the cell is sandwiched between two pieces of polaroid whose polarizing directions correspond to the director ordering direction of the particular cell surfaces they are next to. In the reflective mode a reflector is placed behind the back sheet of polaroid. Figure 4.26 shows the arrangement and traces the behaviour of a polarized beam as it traverses the system. With no applied voltage, the incident light is first polarized, then has its polarization direction rotated by 90° as it traverses the cell, then passes through the second polarizer and is then reflected back along its path where the same process is repeated. With no field applied, therefore, the device reflects incident radiation and appears bright. When a field is applied the direction of polarization of light traversing the cell is not rotated and hence cannot pass through the second polarizer. Little light will then be reflected from the device and it will appear dark.

The amount of light reflected from an LCD as a function of applied voltage is shown schematically in Fig. 4.27. The reflectance, initially constant, falls rapidly beyond a criti-

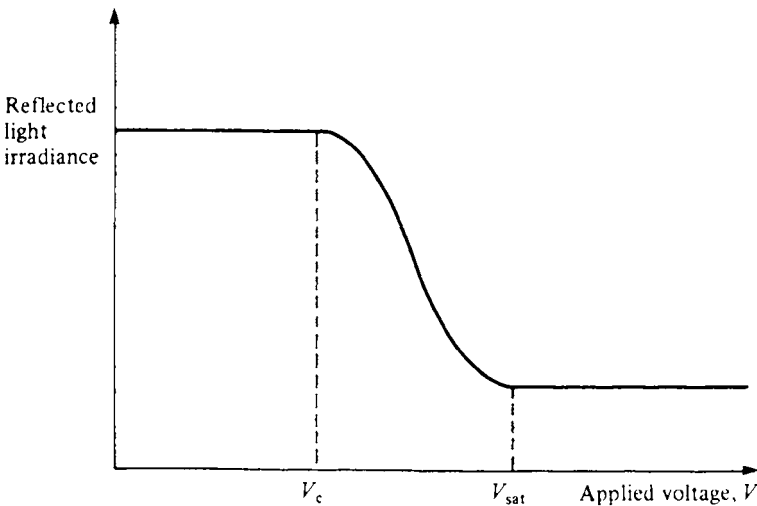


FIG. 4.27 Amount of light reflected from a liquid crystal display as a function of applied voltage V . The reflected irradiance remains constant up to the critical voltage V_c ; it then falls with increasing voltage until it again becomes constant beyond V_{sat} .

cal voltage $V_c (= \epsilon_c D; D = \text{cell thickness})$ and again becomes constant beyond a voltage V_{sat} . A typical value for V_{sat} is 3 V. D.C. operation tends to shorten the operating lifetime of the device owing to electromechanical reactions taking place, and hence a.c. waveforms are invariably used. The cell responds to the r.m.s. value of the voltage waveform. A square waveform which has a frequency of between 25 Hz and 1 kHz is often used.

Transmission LCD displays do not have the reflector, and must be provided with rear illumination, but otherwise they operate in a very similar fashion to the reflective displays. Colour displays are possible by incorporating a colour filter. The use of polarizers in the twisted nematic cell substantially reduces the maximum amount of light that can be reflected from it (see Problem 4.7). In addition, the angle of viewing is found to be restricted to about $\pm 45^\circ$. A greater image contrast over a wider range of viewing angles can be obtained by increasing the angle of twist to 270° to give the so-called 'supertwist' display. The light reflection curve shows a much more rapid switch from on to off states (i.e. the corresponding V_c and V_{sat} values are much closer together than for the 90° twist curve (Fig. 4.27)). A disadvantage of this is that the switching times are longer than for the 90° twist and the cell width has to be reduced to compensate.

4.10 Numeric displays

We turn now to the problem of combining our individual display elements into some pattern capable of conveying more information than just a simple 'on' or 'off' situation. There is no doubt that for medium area, high resolution displays there is as yet no ready alternative to CRTs. These do, however, have the obvious disadvantages of a small display area to volume ratio, of requiring a high voltage supply and of being sensitive to adverse environmental conditions (e.g. stray magnetic fields).

Displays that require only a small number of basic elements are easily catered for by using LED, liquid crystal or plasma display elements. One of the simplest display formats commonly used to form the numbers 0 to 9 consists of seven bar segments and is illustrated in Fig. 4.28(a). Each bar might itself consist of several discrete display elements depending

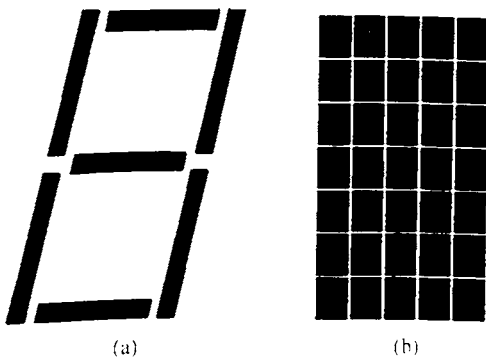


FIG. 4.28 Two common display formats: (a) illustrates the seven-bar segment display used to form the numbers from 0 to 9; the 7×5 matrix display (b) may be used for more complex characters.

on its size. More complex characters may be obtained using a 7×5 matrix (Fig. 4.28b). A display of this latter type can be made either by assembling 35 discrete elements or, in the case of LEDs, a monolithic device can be constructed with all the elements grown onto a single substrate. This latter alternative is only feasible for smaller displays where the characters are less than 5 mm or so in size.

There are two basic methods of wiring up a matrix display. The simpler of these is known as the common anode (or cathode) and is shown in Fig. 4.29(a). The anodes (or cathodes) are all wired together whilst each cathode (or anode) has a separate connection made to it. Thus for N elements the number of external connections is $N + 1$. A technique requiring fewer connections is the *coordinate-connected* (or *matrix-connected*) method shown in Fig. 4.29(b). In this, all the anodes of the elements in each column (row) are connected together as are all the cathodes of the elements in each row (column). For a square array of N elements we thus require $N^{1/2}$ external connections. If N is large, this number is considerably less than that required for the common anode method. It is obvious, however, that a general display pattern cannot be maintained merely by applying static voltages to the leads. The array must be *scanned*. Thus if we label the columns (anodes) by x_1, x_2, x_3 , etc., and the rows (cathodes) by y_1, y_2, y_3 , etc., then a possible scanning sequence is as follows: a voltage $+V$ is applied to column x_1 and zero volts to the rest. Voltages of either $-V$ or $+V$ are then applied

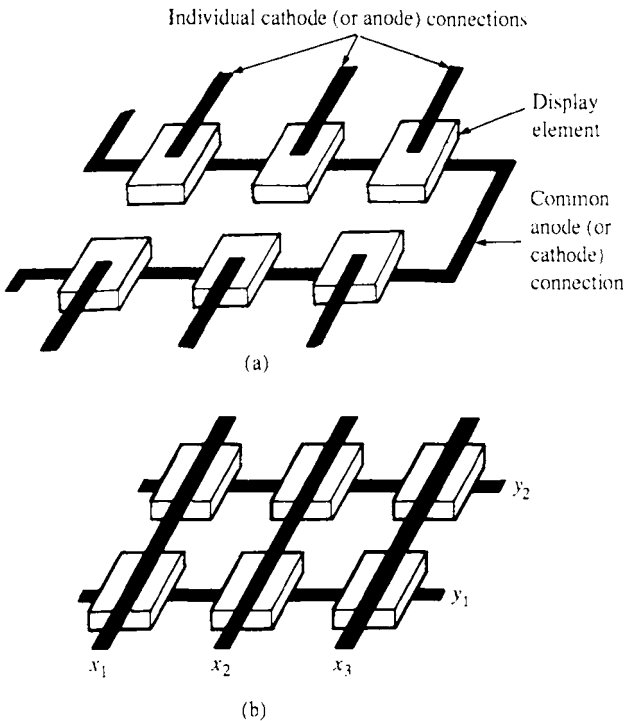


FIG. 4.29 Common anode (or cathode) method of connecting an array of display elements (a). For N elements there will be $N + 1$ connections. (b) The coordinate-connected method of wiring up an array of display elements. The columns are labelled x_1, x_2, x_3 , etc., and the rows y_1, y_2, y_3 , etc.

simultaneously to the rows y_1, y_2, y_3 , etc., the choice depending on whether we require the element at the intersection of column x_1 and the particular row to be on or off. After a time t_c the voltage $+V$ is switched to column x_2 and a new set of voltages applied to the rows.

If the number of columns is N_c then the picture will be scanned in a time $N_c t_c$. To avoid flicker, we must have $N_c t_c \leq 1/45$ s and hence $t_c \leq 1/(45N_c)$ s. If N_c is very large, t_c may approach the response time of the device which will result in a reduced device output. Because each element in the display is on for at most a fraction $1/N_c$ of the total 'on' time of the display then the active display elements must be overdriven by a factor N_c to maintain the same average brightness levels possible from an equivalent common anode display. This places a limitation on the size of N_c unless a much reduced brightness is acceptable.

It will be noticed that even when an element is 'off' it still may have a voltage V across it. An 'ideal' element should thus show no output up to a threshold voltage V_{th} and then give its maximum saturated output at a voltage V_{sat} , where $V_{sat} = 2 \times V_{th}$ (see Fig. 4.30). Several devices do have characteristics that are not too far removed from this ideal, for example the LED and d.c. electroluminescent cells. Often, however, V_{sat} is greater than $2V_{th}$ and so it is not possible to drive the device to saturation, thus reducing the maximum brightness available. In addition, V_{th} may vary with operating temperature and a 'safe' value somewhat below the actual value may have to be adopted, again leading to a reduced output.

Although LCDs have the attraction of requiring very little power, they have so far proved rather unsuitable for matrix displays. They have rather ill-defined threshold voltages which are quite strongly dependent on the operating temperature. In addition, their slow response times prevent them from responding fully to short voltage pulses. Nevertheless it is possible that new materials will be developed which will in part overcome some of these difficulties. (For example, the positive cholesteric materials mentioned in the previous section have a faster response time than the more usual nematic types.)

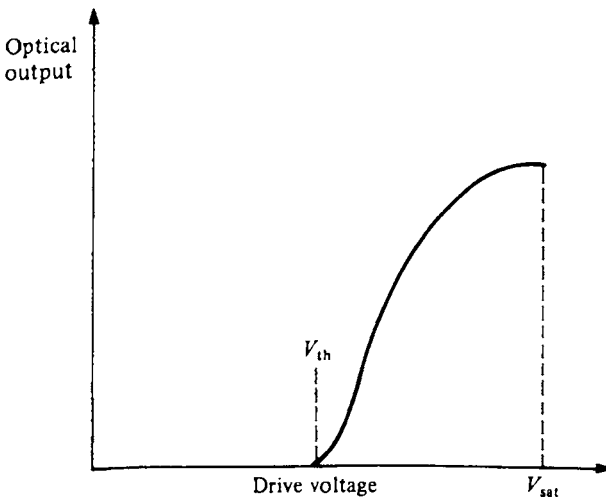


FIG. 4.30 Idealized characteristic for a device suitable for matrix-addressed displays. The device shows no optical output up to a threshold drive voltage V_{th} . Thereafter the output increases to a maximum at a drive voltage of V_{sat} .

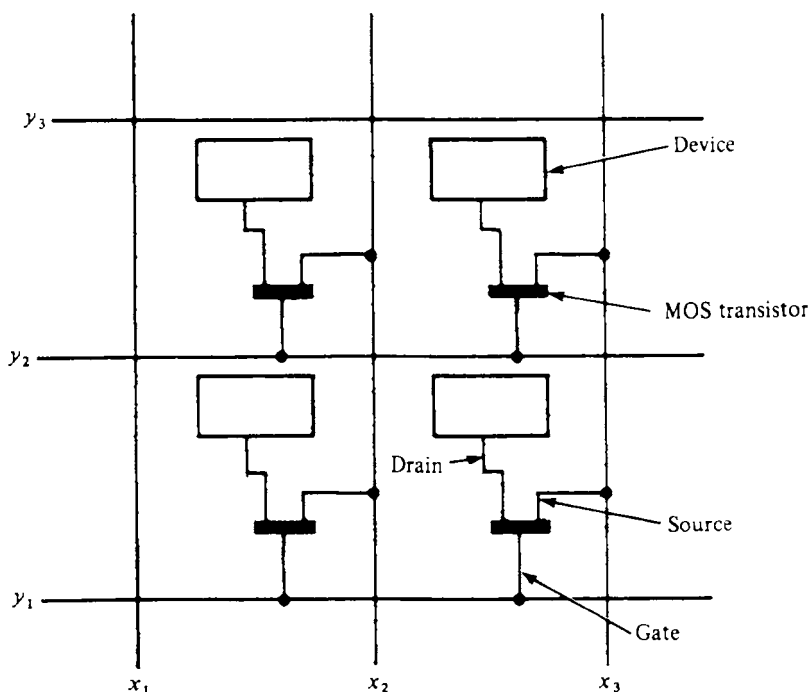


FIG. 4.31 Method used for LCD matrix operation using an MOS transistor. The element may be held 'on' (or 'off') during the time between addressing by virtue of the charge held on the drain terminal. (The LCD consumes minimal power.) To address an element the gate of the attached MOST is 'opened' by applying a suitable potential to the appropriate y line. The voltage that is on the corresponding x line then appears across the device. Note: only one connection to each device is shown here; there will be a further connection common to all elements.

It is possible, when using an LCD, to hold each element on for a time longer than the element address time. This may be achieved using the circuitry shown in Fig. 4.31. Here each element is connected to the drain of an MOS transistor¹ whose source and gate are connected to the appropriate column and row respectively. If the element is to be 'on' then the voltage on the corresponding column is switched onto the element by applying a voltage pulse to the appropriate row. Since the LCD element consumes very little current, the drain capacitance of the MOST may enable the element to remain 'on' during the rest of the scan time when the MOST is switched 'off'. There is, however, the obvious difficulty of reliably fabricating the required number of MOS transistors if N is large.

NOTES

1. For those unfamiliar with the characteristics of an MOS transistor, further details may be found in any of the many books available on solid state devices, for example refs 2.1a and d.

PROBLEMS

- 4.1 Assuming an isotropic radiation source within a transparent medium of refractive index n_1 , show that the fraction F of the radiation that can escape through a plane boundary into another medium of refractive index n_2 ($n_2 < n_1$) is given approximately by

$$F \approx \frac{1}{4} \left(\frac{n_2}{n_1} \right)^2 \left[1 - \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \right]$$

(Hint: assume that when radiation is incident on the boundary at an angle θ (less than the critical angle), then the transmittance $T(\theta)$ is identical with that obtained at normal incidence, i.e. $T(\theta) = T(0) = 1 - [(n_1 - n_2)/(n_1 + n_2)]^2$.)

- 4.2 A GaAs LED fabricated from fairly lightly doped materials has an effective recombination region of width $0.1 \mu\text{m}$. If it is operated at a current density of $2 \times 10^7 \text{ A m}^{-2}$ estimate the modulation bandwidth that can be expected.
- 4.3 Design a simple power supply using a 9 V battery to drive an LED which has a maximum forward current of 20 mA at 1.9 V.
- 4.4 Estimate the average luminance of a CRT screen of area 0.05 m^2 when it is operated with a beam current of $1 \mu\text{A}$ and an accelerating potential of 15 kV. Assume the phosphor used has a power efficiency of 10% and peak emission at a wavelength of 500 nm.
- 4.5 A light source of area A has a luminance at normal incidence of $B(0)$ nits. Calculate the total flux from the source if (a) it is isotropic and (b) it is Lambertian.
- 4.6 An isotropic source is embedded in a transparent medium of refractive index n_1 . Show that when the radiation passes into a second medium of refractive index n_2 , then the source appears to be approximately Lambertian. You may make the same assumptions concerning the transmission of radiation through the interface as are made in Problem 4.1.
- 4.7 Unpolarized light is incident normally onto an LCD cell; if there is no applied voltage what is the maximum fraction of the incident light that can be reflected back?

REFERENCES

- 4.1 G. K. Woodgate, *Elementary Atomic Structure*, McGraw-Hill, London, 1970, Chapter 3.
- 4.2 D. Curie, *Luminescence in Crystals*, Methuen, London, 1960, Chapter VI.
- 4.3 R. C. Alig and S. Bloom, 'Electron-hole pair creation energies in semiconductors', *Phys. Rev. Lett.*, **35**, 1522, 1975.
- 4.4 T. E. Everhart and P. H. Hoff, 'Determination of kilovolt electron energy dissipation vs. distance in solid materials', *J. Appl. Phys.*, **42**, 5837, 1971.
- 4.5 S. Sherr, *Electronic Displays*, John Wiley, New York, 1979, Chapter 2.

- 4.6 H. K. Henisch, *Electroluminescence*, Pergamon Press, Oxford, 1962, Section 1.3.1.
- 4.7 V. P. Varshi, 'Band-to-band radiative recombination in semiconductors', *Phys. Status Solidi*, **19**, 459, 1967.
- 4.8 (a) K. Gillessen and W. Schairer, *Light-Emitting Diodes, an Introduction*, Prentice Hall International, London, 1987, Chapter 3.
(b) A. A. Bergh and P. J. Dean, *Light-Emitting Diodes*, Oxford University Press, Oxford, 1976, Chapter 5.
- 4.9 *Ibid.*, Chapter 1.
- 4.10 S. Chandrasekhar, *Liquid Crystals*, Cambridge University Press, Cambridge, 1980.

Lasers I

The word 'laser' is an acronym for '*light amplification by stimulated emission of radiation*'. Albert Einstein in 1917 showed that the process of stimulated emission must exist, but it was not until 1960 that T. H. Maiman (ref. 5.1) first achieved laser action at optical frequencies in ruby. The basic principles and construction of a laser are relatively straightforward and it is somewhat surprising that the invention of the laser was so long delayed. A rigorous analysis of the physics of the laser, on the other hand, is quite difficult and the treatment we give below is somewhat simplified. The development of lasers since 1960 has been extremely rapid and although applications for lasers had a very slow start during their first decade, new applications for laser radiation are being found now almost every day (see section 6.5); a selection of texts on laser theory and applications is given in ref 5.2. In view of the increased use of lasers, and as laser radiation is potentially hazardous, some comments on laser safety have been included as Appendix 7.

5.1 Emission and absorption of radiation

It was seen in Chapter 1 that when an electron in an atom undergoes transitions between two energy states or levels it either emits or absorbs a photon, which can be described in terms of a wave of frequency ν where $\nu = \Delta E/h$, ΔE being the energy difference between the two levels concerned. Let us consider the electron transitions which may occur between the two energy levels of the hypothetical atomic system shown in Fig. 5.1. If the electron

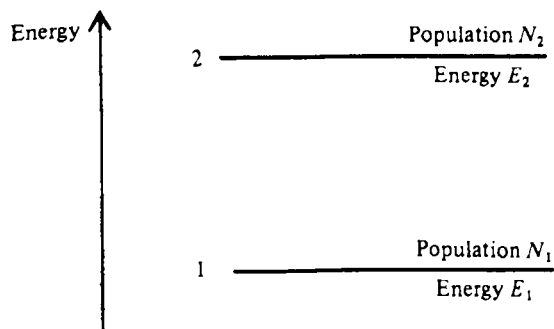


FIG. 5.1 Two-energy-level system.

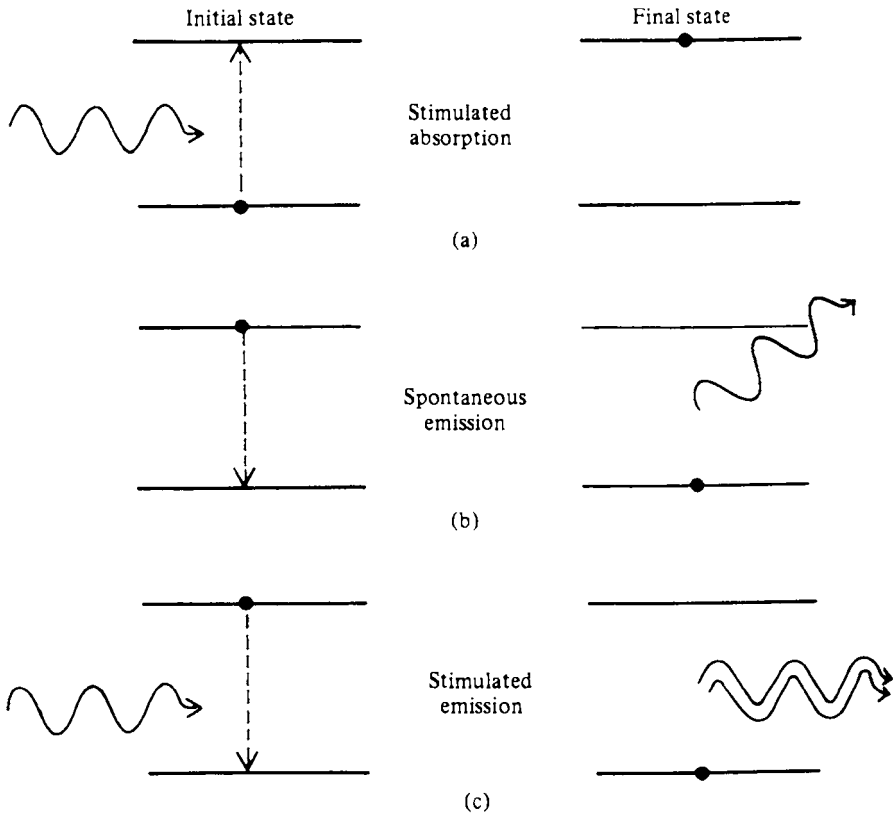


FIG. 5.2 Energy level diagram illustrating (a) absorption, (b) spontaneous emission and (c) stimulated emission. The black dot indicates the state of the atom before and after the transition.

is in the lower level E_1 then in the presence of photons of energy $(E_2 - E_1)$ it may be excited to the upper level E_2 by absorbing a photon. Alternatively if the electron is in the level E_2 it may return to the ground state with the emission of a photon. The emission process may occur in two distinct ways. These are (a) the *spontaneous emission* process in which the electron drops to the lower level in an entirely random way and (b) the *stimulated emission* process in which the electron is 'triggered' to undergo the transition by the presence of photons of energy $(E_2 - E_1)$. There is nothing mystical in this, as the electron would undergo this process sooner or later spontaneously; the transition is simply initiated by the presence of the stimulating photon.

The absorption and emission processes are illustrated in Figs 5.2(a), (b) and (c). Under normal circumstances we do not observe the stimulated process because the probability of the spontaneous process occurring is much higher. The average time the electron exists in the excited state before making a spontaneous transition is called the lifetime τ_{21} of the excited state. The '21' here indicates the energy levels involved. The probability that a particular atom will undergo spontaneous emission within a time interval dt is given by

$A_{21} dt = dt/\tau_{21}$, where A_{21} is the spontaneous transition rate. Because the spontaneous radiation from any atom is emitted at random, the radiation emitted by a large number of atoms will clearly be incoherent. In contrast to this, the stimulated emission process results in coherent radiation since the waves associated with the stimulating and stimulated photons have identical frequencies (but see section 5.7), are in phase, have the same state of polarization and travel in the same direction. This means that with stimulated emission the amplitude of an incident wave can grow as it passes through a collection of excited atoms in what is clearly an amplification process. As the absorption transition, in common with stimulated emission, can only occur in the presence of photons of appropriate energy, it is often referred to as stimulated absorption. These two processes may be regarded as the inverse of one another.

The above discussion ignores the fact that the emission and absorption processes do not simply involve photons of a precisely defined energy. Consequently there is a range of frequencies associated with these processes. The distribution of energies or frequencies in a given transition is described by the *lineshape function* $g(\nu)$, which is discussed in section 5.7.

5.2

Einstein relations

Einstein (ref. 5.3) showed that the parameters describing the above three processes are related through the requirement that for a system in thermal equilibrium the rate of upward transitions (from E_1 to E_2) must equal the rate of the downward transition processes (from E_2 to E_1). Let us suppose that our simple two-level atomic energy level system is in equilibrium inside a blackbody cavity, where, as we saw in section 1.4, the radiation covers a very wide frequency range. Although we indicated at the end of the previous section that the upward and downward transitions involve a spread of frequencies, we may assume that this will be very small compared with that of a blackbody. Nevertheless in further considering the transitions we must take into account the behaviour of photons within these frequency distributions. This analysis is undertaken in Appendix 4; meanwhile we adopt a simpler, approximate approach.

If there are N_1 atoms per unit volume in the collection with energy E_1 , then the upward transition or absorption rate will be proportional to both N_1 and to the number of photons available at the correct frequency. Now ρ_ν , the energy density at frequency ν , is given by $\rho_\nu = N_\nu h\nu$ where N_ν is the number of photons per unit volume having frequency ν . Therefore we may write the upward transition rate as $N_1 \rho_\nu B_{12}$ where B_{12} is a constant. Similarly if there are N_2 atoms per unit volume in the collection with energy E_2 then the induced transition rate from level 2 to level 1 is $N_2 \rho_\nu B_{21}$, where again B_{21} is a constant. The spontaneous transition rate from level 2 to level 1 is simply $N_2 A_{21}$. The total downward transition rate is the sum of the induced and spontaneous contributions, that is

$$N_2 \rho_\nu B_{21} + N_2 A_{21}$$

A_{21} , B_{21} and B_{12} are called the *Einstein coefficients*; the relationships between them can be established as follows.

For a system in equilibrium, the upward and downward transition rates must be equal and

hence we have

$$N_1 \rho_\nu B_{12} = N_2 \rho_\nu B_{21} + N_2 A_{21} \quad (5.1)$$

Thus

$$\rho_\nu = \frac{N_2 A_{21}}{N_1 B_{12} - N_2 B_{21}}$$

or

$$\rho_\nu = \frac{A_{21}/B_{21}}{(B_{12}/B_{21})(N_1/N_2) - 1} \quad (5.2)$$

Now, the populations of the various energy levels of a system in thermal equilibrium are given by Boltzmann statistics to be

$$N_j = \frac{g_j N_0 \exp(-E_j/kT)}{\sum g_i \exp(-E_i/kT)} \quad (5.3)$$

where N_j is the population density of the energy level E_j , N_0 is the total population density and g_j is the degeneracy of the j th level.¹ Hence

$$\frac{N_1}{N_2} = \frac{g_1}{g_2} \exp[(E_2 - E_1)/kT] = \frac{g_1}{g_2} \exp(h\nu/kT) \quad (5.4)$$

Therefore, substituting eq. (5.4) into eq. (5.2) gives

$$\rho_\nu = \frac{A_{21}/B_{21}}{[(g_1/g_2)(B_{12}/B_{21}) \exp(h\nu/kT)] - 1} \quad (5.5)$$

Since the collection of atoms in the system we are considering is in thermal equilibrium it must give rise to radiation which is identical with blackbody radiation, eq. (1.42), the radiation density of which can be described by (see ref. 1.1c)

$$\rho_\nu = \frac{8\pi h\nu^3 n^3}{c^3} \left(\frac{1}{\exp(h\nu/kT) - 1} \right) \quad (5.6)$$

where n is the refractive index of the medium.

Comparing eqs (5.5) and (5.6) for ρ_ν , we see that

$$g_1 B_{12} = g_2 B_{21} \quad (5.7)$$

and

$$\frac{A_{21}}{B_{21}} = \frac{8\pi h\nu^3 n^3}{c^3} \quad (5.8)$$

Equations (5.7) and (5.8) are referred to as the *Einstein relations*. The second relation enables us to evaluate the ratio of the rate of spontaneous emission to the rate of stimulated emission for a given pair of energy levels. We see that this ratio is given by

$$R = \frac{A_{21}}{\rho_\nu B_{21}} \quad (5.9)$$

or

$$R = \exp(h\nu/kT) - 1 \quad (5.9a)$$

EXAMPLE 5.1 Ratio of rates of spontaneous and stimulated emission

Let us calculate this ratio for a tungsten filament lamp operating at a temperature of 2000 K. Taking the average frequency to be 5×10^{14} Hz we see from eq. (5.9b) that the ratio is

$$R = \exp\left(\frac{6.6 \times 10^{-34} \times 5 \times 10^{14}}{1.38 \times 10^{-23} \times 2000}\right) - 1 \approx \exp(12) \approx 1.5 \times 10^5$$

This confirms that under conditions of thermal equilibrium, stimulated emission is not an important process. For sources operating at lower temperatures and higher frequencies, stimulated emission is even less likely.

The above discussion indicates that the process of stimulated emission competes with the processes of spontaneous emission and absorption. Clearly, if we wish to amplify a beam of light by stimulated emission then we must increase the rate of this process in relation to the other two processes. Consideration of eq. (5.1) indicates that to achieve this for a given pair of energy levels we must increase both the radiation density and the population density N_2 of the upper level in relation to the population density N_1 of the lower level. Indeed, we shall show that to produce laser action we must create a condition in which $N_2 > (g_2/g_1)N_1$ even though $E_2 > E_1$, that is we must create a so-called *population inversion*. Before describing this situation in detail it will be instructive to look more closely at the process of absorption.

5.3 Absorption of radiation

Let us consider a collimated beam of perfectly monochromatic radiation of unit cross-sectional area passing through an absorbing medium. We assume for simplicity that there is only one relevant electron transition, which occurs between the energy levels E_1 and E_2 . Then the change in irradiance of the beam as a function of distance is given by

$$\Delta I(x) = I(x + \Delta x) - I(x)$$

For a homogeneous medium $\Delta I(x)$ is proportional both to the distance travelled Δx and to $I(x)$. That is, $\Delta I(x) = -\alpha I(x)\Delta x$, where the constant of proportionality, α , is the *absorption coefficient*. The negative sign indicates the reduction in beam irradiance due to absorption as α is a positive quantity. Writing this expression as a differential equation we have

$$\frac{dI(x)}{dx} = -\alpha I(x)$$

Integrating this equation gives

$$I = I_0 \exp(-\alpha x) \tag{5.10}$$

where I_0 is the incident irradiance.

Let us consider the absorption coefficient in more detail. Clearly, the degree of absorption of the beam will depend on how many atoms N_1 there are with electrons in the lower energy state E_1 and on how many atoms N_2 there are in energy state E_2 . If N_2 were zero, then absorption would be a maximum, while if all of the atoms were in the upper state the absorption would be zero and the probability of stimulated emission would be large.

From the discussion of induced or stimulated transitions given in section 5.2, we can write an expression for the net rate of loss of photons per unit volume, $-dN_v/dt$, from the beam as it travels through a volume element of medium of thickness Δx and unit cross-sectional area (Fig. 5.3) as

$$-\frac{dN_v}{dt} = N_1 \rho_v B_{12} - N_2 \rho_v B_{21}$$

or substituting from eq. (5.7)

$$-\frac{dN_v}{dt} = \left(\frac{g_2}{g_1} N_1 - N_2 \right) \rho_v B_{21} \tag{5.11}$$

In this discussion, we have deliberately ignored photons generated by spontaneous emission as these are emitted randomly in all directions and do not therefore contribute to the collimated beam. Similarly we have ignored scattering losses.

We can now link eq. (5.11), which contains the difference in populations of the two energy levels, to the absorption coefficient α . We recall that the irradiance of the beam is the energy crossing unit area in unit time and therefore is given by the energy density times the speed of light in the medium, that is $I = \rho c/n$, or for photons of frequency ν , $I_v = \rho_v c/n = N_v h\nu c/n$, where c is the speed of light in *vacuo* and n is the refractive index of the medium. Hence the change in photon density within the beam between the boundaries x and $x + \Delta x$ of the

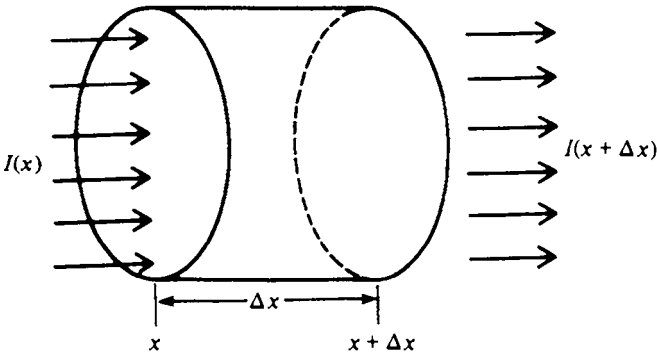


FIG. 5.3 Radiation passing through a volume element of length Δx and unit cross-sectional area.

volume element can be written as

$$-d\mathcal{N}_v(x) = [I_v(x) - I_v(x + \Delta x)] \frac{n}{h\nu_{21}c}$$

If Δx is sufficiently small we can rewrite this equation as

$$-d\mathcal{N}_v(x) = -\frac{dI_v(x)}{dx} \cdot \frac{\Delta x n}{h\nu_{21}c}$$

Thus the rate of decay of photon density in a time interval $dt (= \Delta x/(c/n))$ is

$$\frac{d\mathcal{N}_v}{dt} = \frac{dI_v(x)}{dx} \cdot \frac{1}{h\nu_{21}}$$

and substituting for dI/dx from eq. (5.10) gives

$$\frac{d\mathcal{N}_v}{dt} = -\alpha I_v(x) \cdot \frac{1}{h\nu_{21}} = -\alpha \rho_v \frac{c}{n} \cdot \frac{1}{h\nu_{21}} \quad (5.12)$$

Hence comparing eqs (5.11) and (5.12) we have

$$\alpha \rho_v \frac{c}{n} \frac{1}{h\nu_{21}} = \left(\frac{g_2}{g_1} N_1 - N_2 \right) \rho_v B_{21}$$

Therefore the absorption coefficient α is given by

$$\alpha = \left(\frac{g_2}{g_1} N_1 - N_2 \right) \frac{B_{21} h\nu_{21} n}{c} \quad (5.13)$$

We see from eq. (5.13) that α , as we supposed earlier, depends on the difference in the populations of the two energy levels E_1 and E_2 . For a collection of atoms in thermal equilibrium, since $E_2 > E_1$, $(g_2/g_1)N_1$ will always be greater than N_2 (eq. 5.4) and hence α is positive. If, however, we can create a situation, referred to in the previous section, in which N_2 becomes greater than $(g_2/g_1)N_1$ then α is negative and the quantity $(-\alpha x)$ in the exponent of eq. (5.10) becomes positive. Thus the irradiance of the beam grows as it propagates through the medium in accordance with the equation

$$I = I_0 \exp(kx) \quad (5.14)$$

where k is referred to as the *small signal gain coefficient* and is given by

$$k = \left(N_2 - \frac{g_2}{g_1} N_1 \right) B_{21} \frac{h\nu_{21} n}{c} \quad (5.15)$$

5.4

Population inversion

The population inversion condition required for light amplification is a non-equilibrium distribution of atoms among the various energy levels of the atomic system. The Boltzmann

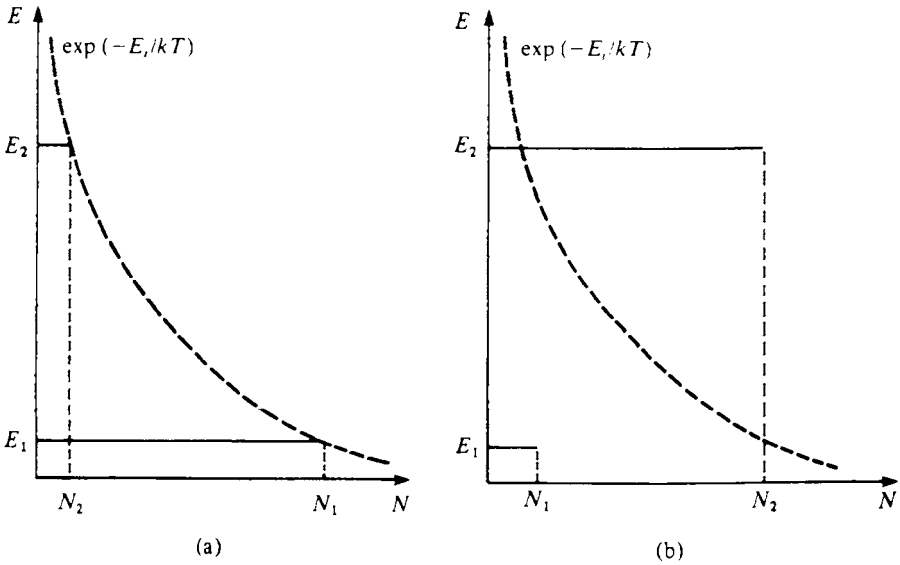


FIG. 5.4 Populations of a two-level energy system: (a) in thermal equilibrium; (b) after a population inversion has been produced.

distribution which applies to a system in thermal equilibrium is given by eq. (5.3) and is illustrated in Fig. 5.4(a); N_j is the population density of the j th energy level and clearly as E_j increases N_j decreases for a constant temperature. We note that if the energy difference between E_1 and E_2 were nearly equal to kT (≈ 0.025 eV at room temperature) then the population of the upper level would be $\exp(-1)$ or 0.37 of that of the lower level. For an energy difference large enough to give visible radiation (≈ 2.0 eV), however, the population of the upper level is almost negligible as Example 5.2 shows.

EXAMPLE 5.2 Relative populations of energy levels

We may estimate the relative populations in thermal equilibrium of two energy levels such that a transition from the higher to the lower will give visible radiation.

Let us take the average wavelength of visible radiation as 550 nm; then $E_2 - E_1 = hc/(550 \times 10^{-9}) = 3.6 \times 10^{-19} \text{ J} = 2.25 \text{ eV}$.

Assuming room temperature ($T \approx 300 \text{ K}$) and that $g_1 = g_2$, we have from eq. (5.4)

$$\frac{N_2}{N_1} = \exp\left(\frac{-3.6 \times 10^{-19}}{1.38 \times 10^{-23} \times 300}\right) \approx \exp(-87) \approx 10^{-37}$$

Clearly, then, if we are to create a population inversion, illustrated in Fig. 5.4(b), we must supply a large amount of energy to excite atoms into the upper level E_2 . This excitation process is called *pumping* and much of the technology of lasers is concerned with how the

pumping energy can be supplied to a given laser system. Pumping produces a non-thermal equilibrium situation; we shall now consider in general terms how pumping enables a population inversion to be achieved.

5.4.1 Attainment of a population inversion

One of the methods used for pumping is stimulated absorption: that is, the energy levels which one hopes to use for laser action are pumped by intense irradiation of the system. Now as B_{12} and B_{21} are equal (assuming $g_1 = g_2$), once atoms are excited into the upper level the probabilities of further stimulated absorption or emission are equal so that even with very intense pumping the best that can be achieved with the two-level system, considered hitherto, is equality of the populations of the two levels.

As a consequence we must look for materials with either three or four energy level systems; this is not really a disadvantage as atomic systems generally have a large number of energy levels.

The three-level system first proposed by Bloembergen (ref. 5.4) is illustrated in Fig. 5.5. Initially the distribution obeys Boltzmann's law. If the collection of atoms is intensely illuminated the electrons can be excited (i.e. pumped) into the level E_2 from the ground state E_0 . From E_2 the electrons decay by non-radiative processes to the level E_1 and a population inversion may be created between E_1 and E_0 . Ideally, the transition from level E_2 to E_1 should be very rapid, thereby ensuring that there are always vacant states at E_2 , while that from E_1 to E_0 should be very slow, that is E_1 should be a *metastable* state. This allows a large build-up in the number of atoms in level E_1 , as the probability of spontaneous emission is relatively small. Eventually N_1 may become greater than N_0 and then population inversion will have been achieved.

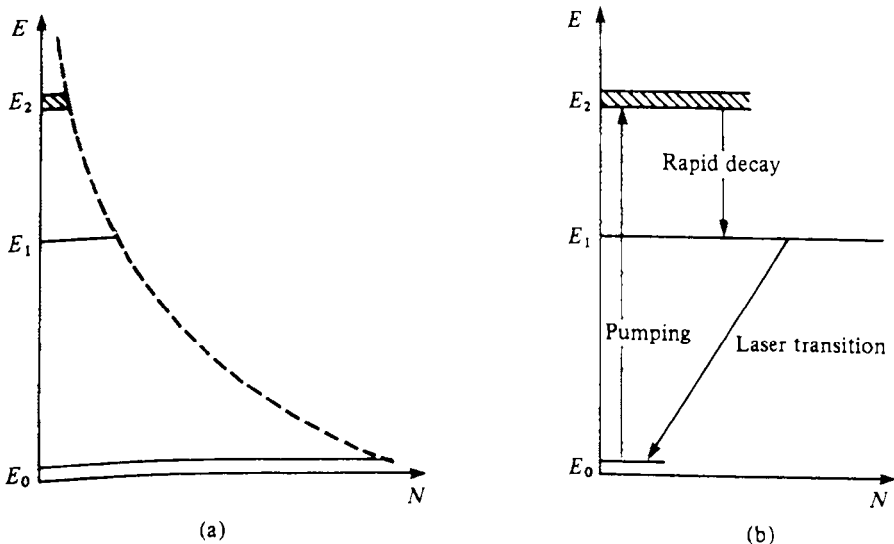


FIG. 5.5 Population of the energy levels by pumping in a three-level system: (a) Boltzmann distribution before pumping and (b) distribution after pumping and the transitions involved.

The level E_2 should preferably consist of a large number of closely spaced levels so that pumping uses as wide a part of the spectral range of the pumping radiation as possible, thereby increasing the pumping efficiency. Even so, three-level lasers, for example ruby, require very high pump powers because the terminal level of the laser transition is the ground state. This means that rather more than half of the ground state atoms (this number is usually very nearly equal to the total number of atoms in the collection) have to be pumped to the upper state to achieve a population inversion.

The four-level system shown in Fig. 5.6 has much lower pumping requirements. If $(E_1 - E_0)$ is rather large compared with kT (the thermal energy at the temperature of operation), then the populations of the levels E_1 , E_2 and E_3 are all very small in conditions of thermal equilibrium. Thus, if atoms are pumped from the ground state to the level E_3 from which they decay very rapidly to the metastable level E_2 , a population inversion is quickly created between levels E_2 and E_1 .

Again the upper level E_3 should preferably consist of a large number of levels for greatest pumping efficiency. If the lifetimes of the transitions $E_3 \rightarrow E_2$ and $E_1 \rightarrow E_0$ are short, the population inversion between E_2 and E_1 can be maintained with moderate pumping and continuous laser action can be achieved more readily. In the Nd:YAG laser, for example, $\tau_{21} \approx 0.5$ ms while $\tau_{10} \approx 30$ ns and, although there are many upper levels used for pumping, each has a lifetime of about 10^{-8} s (i.e. $\tau_{32} \approx 10^{-8}$ s). The details of the mechanisms used for pumping lasers can be quite complicated and, in addition to optical pumping, pumping can occur in an electrical discharge or by electron bombardment, the release of chemical energy, the passage of a current, etc. The energy level schemes of the media used in lasers are often complex, but they can usually be approximated by either three- or four-level schemes.

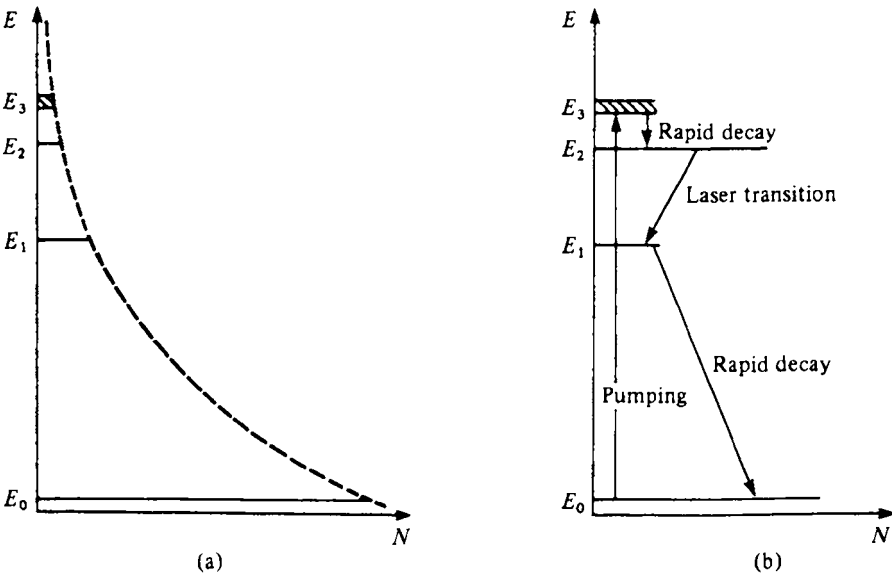


FIG. 5.6 Population of the energy levels in a four-level system: (a) before pumping and (b) after pumping.

5.5

Optical feedback

The laser, despite its name, is more analogous to an oscillator than an amplifier. In an electronic oscillator, an amplifier tuned to a particular frequency is provided with positive feedback and, when switched on, any electrical noise signal of the appropriate frequency appearing at the input will be amplified. The amplified output is fed back to the input and amplified yet again and so on. A stable output is quickly reached, however, since the amplifier saturates at high input voltages, as it cannot produce a larger output than the supply voltage.

In the laser, positive feedback may be obtained by placing the gain medium between a pair of mirrors which, in fact, form an optical cavity (a Fabry – Perot resonator). The initial stimulus is provided by any spontaneous transitions between appropriate energy levels in which the emitted photon travels along the axis of the system. The signal is amplified as it passes through the medium and ‘fed back’ by the mirrors. Saturation is reached when the gain provided by the medium exactly matches the losses incurred during a complete round trip.

The gain per unit length of many active media is so small that very little amplification of a beam of light results from a single pass through the medium. In the multiple passes which a beam undergoes when the medium is placed within a cavity, however, the amplification may be substantial.

We have tacitly assumed that the radiation within the cavity propagates to and fro between two plane – parallel mirrors in a well-collimated beam. Because of diffraction effects, however, this cannot be the case as a perfectly collimated beam cannot be maintained with mirrors of finite extent; some radiation will spread out beyond the edges of the mirrors. Diffraction losses of this nature can be reduced by using concave mirrors. In practice a number of different mirror curvatures and configurations are used depending on the applications envisaged and the type of laser being used.

A detailed analysis of the effects of different mirror systems requires a rigorous application of diffraction theory and is beyond the scope of this book (see e.g. ref. 5.2a). Using simple ray tracing techniques, however, it is quite easy to anticipate the results of such an analysis in that mirror configurations which retain a ray of light, initially inclined at a small angle to the axis, within the optical cavity after several reflections are likely to be useful (see ref. 5.5). Such cavities are said to be stable.

The commonly used mirror configurations are shown in Fig. 5.7; they all have various advantages and disadvantages. The plane – parallel configuration, for example, is very difficult to align, for if the mirrors are not strictly parallel (to within about 1 second of arc) the optical beam will ‘walk off’ the mirrors after a few reflections. On the other hand, the radiation beam makes maximum use of the laser medium (we say that it has a large *mode volume* – see also section 5.9) as there is no focusing of the beam within the cavity. In addition, the mirrors need to be flat to within $\lambda/100$. In contrast to the plane – parallel case, the confocal arrangement is relatively easy to align (an accuracy of 1.5 minutes of arc is sufficient) but the use of the active medium is restricted (i.e. the mode volume is small). In gas lasers, if maximum power output is required we use a large radius resonator, while if uniphase operation (i.e. maximum beam coherence) is required we use the hemispherical system.

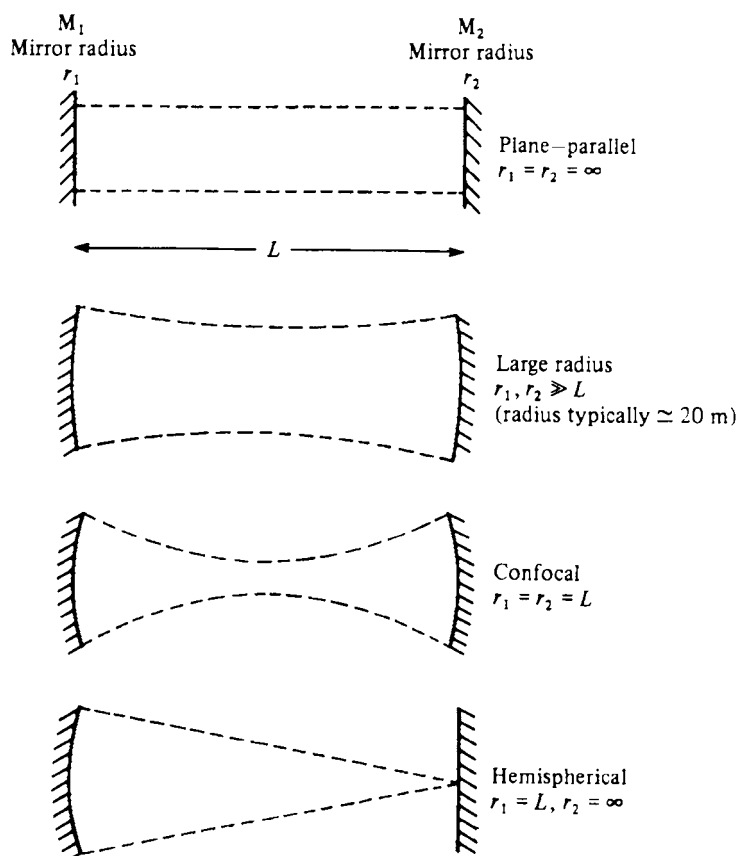


FIG. 5.7 Some commonly used laser cavity mirror configurations (the dashed lines show the extent of the mode volume in each case).

Sometimes mirror configurations are used which give rise to unstable cavities. In these a ray which is initially travelling at a small angle of incidence to the cavity axis will diverge away from the axis after a number of reflections. Such resonators are characterized by high losses, but even so they have some useful properties. In particular they can make efficient use of the mode volume. As unstable resonators have large losses they can be used effectively only with high gain media such as carbon dioxide. As we mentioned earlier, the gain is usually very small so it is essential to minimize all losses in the laser (see section 5.6). One source of loss is absorption in the mirrors. To reduce this, high reflectance multilayer dielectric coatings on the mirrors are used rather than metallic coatings. In these so-called multilayer stacks there is a sequence of quarter-wave (i.e. $\lambda/4$) layers of alternate high and low refractive index dielectric materials on a glass substrate. Because of the phase changes occurring at alternate interfaces, all the reflected waves are in phase and add constructively. More than 20 such layers may be needed to give reflectances in excess of 99.9% – lower reflectances require fewer layers. Clearly, the mirrors will only be effective over a narrow

wavelength range. A familiar example of this sort of process is the blooming of camera lenses to reduce unwanted reflections.

We can now derive the minimum pump power required (i.e. the threshold condition) in terms of the parameters of the whole system for laser oscillations to occur.

5.6

Threshold conditions – laser losses

It was explained above that a steady state level of oscillation is reached when the rate of amplification is balanced by the rate of loss. This is the situation in continuous output (CW) lasers; it is a little different in pulse lasers. Thus, while a population inversion is a necessary condition for laser action, it is not a sufficient one because the minimum (i.e. threshold value) of the gain coefficient must be large enough to overcome the losses and sustain oscillations. The threshold gain, in turn, through eq. (5.15) specifies the minimum population inversion required.

The total loss of the system is due to a number of different processes; the most important ones include:

1. Transmission at the mirrors – the transmission from one of the mirrors usually provides the useful output, the other mirror is made as reflective as possible to minimize losses.
2. Absorption and scattering at the mirrors.
3. Absorption in the laser medium due to transitions other than the desired transitions (as mentioned earlier most laser media have many energy levels, not all of which will be involved in the laser action).
4. Scattering at optical inhomogeneities in the laser medium – this applies particularly to solid state lasers.
5. Diffraction losses at the mirrors.

To simplify matters, let us include all the losses except those due to transmission at the mirrors in a single effective loss coefficient γ which reduces the effective gain coefficient to $(k - \gamma)$. We can determine the threshold gain by considering the change in irradiance of a beam of light undergoing a round trip within the laser cavity. We assume that the laser medium fills the space between the mirrors M_1 and M_2 which have reflectances R_1 and R_2 and a separation L . Then in travelling from M_1 to M_2 the beam irradiance increases from I_0 to I where, from eq. (5.14),

$$I = I_0 \exp[(k - \gamma)L]$$

After reflection at M_2 , the beam irradiance will be $R_2 I_0 \exp[(k - \gamma)L]$ and after a complete round trip the final irradiance will be such that the round trip gain G is

$$G = \frac{\text{final irradiance}}{\text{initial irradiance}} = R_1 R_2 \exp[2(k - \gamma)L]$$

If G is greater than unity a disturbance at the laser resonant frequency will undergo a net amplification and the oscillations will grow; if G is less than unity the oscillations will die

out. Therefore we can write the *threshold condition* as

$$G = R_1 R_2 \exp[2(k_{\text{th}} - \gamma)L] = 1 \quad (5.16)$$

where k_{th} is the threshold gain. It is important to realize that the threshold gain is equal to the steady state gain in continuous output lasers, that is $k_{\text{th}} = k_{\text{ss}}$. This equality is due to a phenomenon known as *gain saturation*, which can be explained as follows. Initially, when laser action commences the gain may be well above the threshold value. The effect of stimulated emission, however, will be to reduce the population of the upper level of the laser transition so that the degree of population inversion and consequently the gain will decrease. Thus the net round trip gain may vary and be greater than or less than unity so that the cavity energy density will correspondingly increase or decrease. It is only when G has been equal to unity for a period of time that the cavity energy (and laser output power) settles down to a steady state value, that is when the gain just balances the losses in the medium. In terms of the population inversion there will be a threshold value $N_{\text{th}} = [N_2 - (g_2/g_1)N_1]_{\text{th}}$ corresponding to k_{th} . In steady state situations $[N_2 - (g_2/g_1)N_1]$ remains equal to N_{th} regardless of the amount by which the threshold pumping rate is exceeded (see section 5.8). The small signal gain required to support steady state operation depends on the laser medium through k and γ , and on the laser construction through R_1 , R_2 and L . From eq. (5.16) we can see that

$$k_{\text{th}} = \gamma + \frac{1}{2L} \ln \left(\frac{1}{R_1 R_2} \right) \quad (5.17)$$

where the first term represents the volume losses and the second the loss in the form of the useful output. Equation (5.15) shows that k can have a wide range of values, depending not only on $[N_2 - (g_2/g_1)N_1]$ but also on the intrinsic properties of the active medium. If k is high then it is relatively easy to achieve laser action, mirror alignment is not critical and dust can be tolerated on the mirrors. With low gain media, on the other hand, such losses are unacceptable and the mirrors must have high reflectances, and be scrupulously clean and carefully aligned.

It should be noted that a laser with a high gain medium will not necessarily have a high efficiency. The efficiency is the ratio of the output light power to the input pumping power. It therefore depends on how effectively the pump power is converted into producing a population inversion, on the probabilities of different kinds of transitions from the upper level and on the losses in the system. With reference to Fig. 5.6(b), and confining our attention to optical pumping, we can easily see that the efficiency cannot exceed $(E_2 - E_1)/(E_3 - E_0) = \nu_{21}/\nu_{30}$ for the four-level system and that it will be considerably less than this for the three-level system, where over half the atoms have to be pumped out of the ground state before population inversion is produced. The actual efficiencies, as defined above, are usually very much less than this because of the energy loss in converting electrical energy into optical energy at the pump frequency and the fact that not all the atoms pumped into level 3 will necessarily make a transition to level 2. Certain lasers (e.g. CO_2) are characterized by having a high efficiency and a high small signal gain. Other lasers such as argon, although having a high gain, have a very low efficiency.

5.7

Lineshape function

In deriving the expression for the small signal gain we assumed that all the atoms in either the upper or lower levels would be able to interact with the (perfectly) monochromatic beam. In fact this is not so; spectral lines have a finite wavelength (or frequency) spread, that is they have a spectral width. This can be seen in both emission and absorption and if, for example, we were to measure the transmission as a function of frequency for the transition between the two energy states E_1 and E_2 , we would obtain the bell-shape curve shown in Fig. 5.8(a).

The emission curve would be the inverse of this (see Fig. 5.8b). The shape of these curves is described by the *lineshape function* $g(\nu)$, which can also be used to describe a frequency probability curve. Thus we may define $g(\nu) d\nu$ as the probability that a given transition between the two energy levels will result in the emission (or absorption) of a photon whose frequency lies between ν and $\nu + d\nu$. $g(\nu)$ is normalized such that $\int_{-\infty}^{\infty} g(\nu) d\nu = 1$. Therefore we see that a photon of energy $h\nu$ may not necessarily stimulate another photon of energy $h\nu$. We then take $g(\nu) d\nu$ as the probability that the stimulated photon will have an energy between $h\nu$ and $h(\nu + d\nu)$.

It is shown in Appendix 4 that, when a monochromatic beam of frequency ν_s interacts with a group of atoms with a lineshape function $g(\nu)$, the small signal gain coefficient may be written as

$$k(\nu_s) = \left(N_2 - \frac{g_2}{g_1} N_1 \right) \frac{B_{21} h \nu_s n g(\nu_s)}{c} \quad (5.18)$$

The form of the lineshape function $g(\nu)$ depends on the particular mechanism responsible

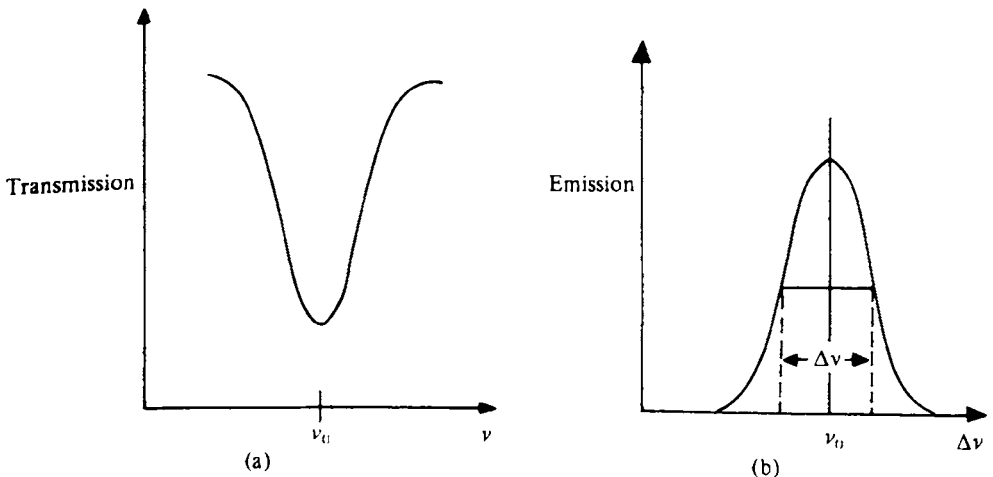


FIG. 5.8 Transmission curve for transitions between energy levels E_1 and E_2 (a) and the emission curve for transitions between E_2 and E_1 (b). The precise form of these curves (the lineshape) depends on the spectral broadening mechanisms.

for the *spectral broadening* in a given transition. The three most important mechanisms are Doppler broadening, collision (or pressure) broadening and natural (or lifetime) damping, which are described briefly below; the interested reader is referred to ref. 5.6 for further details.

EXAMPLE 5.3 Small signal gain coefficient

It may be instructive to calculate the value of population inversion required to give a gain coefficient of 1 m^{-1} in a given laser. We take Nd:YAG for which we have the following data: spontaneous lifetime, $\tau_{21} = 230 \text{ } \mu\text{s}$; wavelength, $\lambda = 1.06 \text{ } \mu\text{m}$; refractive index, $n = 1.82$; and linewidth, $\Delta\nu = 3 \times 10^{12} \text{ Hz}$.

From eq. (5.8) we have $B_{21} = \lambda_0^3 / (8\pi h \tau_{21} n^3) = 5.15 \times 10^{16} \text{ m}^3 \text{ W}^{-1} \text{ s}^{-3}$. Therefore from eq. (5.18) we have (with $k = 1$ and $g(\nu_s) \approx 1/\Delta\nu$, see eq. 5.21)

$$\left(N_2 - \frac{g_2}{g_1} N_1 \right) = \frac{k \lambda_0 \Delta\nu}{B_{21} h n} \approx 5.12 \times 10^{22} \text{ m}^{-3}$$

DOPPLER BROADENING

We are familiar with the Doppler effect, which occurs because of the relative motion of a source and observer. The frequency as measured by the observer increases if the source and observer approach one another and decreases as they recede. This effect applies to a collection of atoms emitting at an optical frequency ν_{12} , so that the observed frequency is given by

$$\nu'_{12} = \nu_{12} \left(1 \pm \frac{v_x}{c} \right)$$

where v_x is the component of the velocity of the atom along the direction of observation (we assume $v_x \ll c$). Since the atoms are in random motion, an observer would measure a range of frequencies depending on the magnitude and direction of v_x . That is, as far as the observer is concerned the collection of atoms would be emitting at a range of different resonant frequencies resulting in a broadening of the emission lineshape. The individual Doppler-shifted resonant frequencies contribute to the smooth Doppler-broadened lineshape.

The mean-squared velocity components v_x depend on the temperature according to $\frac{1}{2} M v_x^2 = \frac{1}{2} kT$, where M is the atomic mass, so that the halfwidth (full width of the curve at half the maximum intensity of emission) of the curve is proportional to the square root of T . Doppler broadening is the predominant mechanism in most gas lasers emitting in the visible region. For example, the halfwidth of the 632.8 nm transition of the He-Ne laser is about $2 \times 10^{-3} \text{ nm}$, assuming a temperature of operation of about 400 K (see Problem 5.6).

Halfwidths are often expressed in terms of frequency: thus a halfwidth of $2 \times 10^{-3} \text{ nm}$ corresponds to a frequency halfwidth of 1500 MHz. (As $c = \nu\lambda$, we may write $d\nu = -(c/\lambda^2) d\lambda$ and hence $d\nu = (3 \times 10^8 \times 2 \times 10^{-12}) / (632.8 \times 10^{-9})^2 = 1500 \text{ MHz}$.)

COLLISION BROADENING

The Doppler linewidth of molecular lasers such as the CO_2 laser is relatively small because of their low resonant frequencies (in the infrared) and comparatively large molecular masses. In such lasers, collision broadening becomes important. Collision broadening also occurs in doped insulator lasers. In these lasers the ions of the active medium may suffer collisions with phonons, that is quantized lattice vibrations.

If an atom which is emitting a photon suffers a collision, then the phase of the wave train associated with the photon is suddenly altered. This in effect shortens the emitted wave trains, which can be shown by Fourier techniques (ref. 5.7) to be equivalent to a broadening of the spectral line. Clearly, the higher the pressure (and temperature) of the gas the more frequently will the atoms suffer collisions and the greater will be the spectral broadening.

NATURAL DAMPING

It can be shown that the very act of an atomic electron emitting energy in the form of a photon leads to an exponential damping of the amplitude of the wave train. The effect of this is similar to collision broadening in that it effectively shortens the wave trains and produces a broadened spectral line.

Broadening mechanisms can be classified into *homogeneous* and *inhomogeneous* broadening. If all of the atoms of the collection have the same transition centre frequency and the same resonance lineshape then the broadening is termed homogeneous; such is the case for collision broadening. On the other hand, in some situations each atom has a slightly different resonance frequency or lineshape for the same transition. The observed lineshape is then the average of the individual ones, such as in Doppler broadening, and the mechanism is termed inhomogeneous. Local variations of temperature, pressure, applied magnetic field as well as local variations due to crystal imperfections also lead to inhomogeneous broadening of the emission or absorption lineshapes. The nature of the broadening is important in several aspect of laser theory, for example in the discussion of gain saturation mentioned earlier.

Homogeneous broadening mechanisms lead to a Lorentzian lineshape which may be written as

$$g(\nu)_L = \frac{\Delta\nu}{2\pi} \left[(\nu - \nu_0)^2 + \left(\frac{\Delta\nu}{2} \right)^2 \right]^{-1}$$

where $\Delta\nu$ is the linewidth, that is the separation between the two points on the (frequency) curve where the function falls to half of its peak value which occurs at frequency ν_0 . Putting $\nu = \nu_0$ gives

$$g(\nu_0)_L = \frac{2}{\pi\Delta\nu} \quad (5.19)$$

Inhomogeneous broadening mechanisms, on the other hand, lead for a gas to a Gaussian frequency distribution, given by

$$g(\nu)_G = \frac{2}{\Delta\nu} \left(\frac{\ln 2}{\pi} \right)^{1/2} \exp \left[-(\ln 2) \left(\frac{\nu - \nu_0}{\Delta\nu/2} \right)^2 \right]$$

and putting $\nu = \nu_0$ gives

$$g(\nu_0)_G = \frac{2}{\Delta\nu} \left(\frac{\ln 2}{\pi} \right)^{1/2} \quad (5.20)$$

For the purpose of later calculations we may, in fact, approximate both eqs (5.19) and (5.20) by

$$g(\nu_0) \approx \frac{1}{\Delta\nu} \quad (5.21)$$

Hence at the frequency $\nu_s = \nu_0$ we may approximate $g(\nu_s)$ by $1/\Delta\nu$ in eq. (5.18) for the small signal gain coefficient.

Because of these various broadening mechanisms, we can no longer treat a group of atoms as though they all radiate at the same frequency. Instead, we must consider a small spread of frequencies about some central value. It might then be expected that the output of the laser would contain the same distribution of frequencies as the broadened transitions of the atoms in the medium. This is, in fact, not the case as the spectral character of the laser output is different from that of spontaneous emission in the same medium. Two factors account for this difference: the effects of the optical resonator (discussed below) and the effect of the amplification process on the irradiance. As light travels through an amplifying medium the irradiance varies as

$$I_\nu(\nu, x) = I(\nu, 0) \exp[k(\nu)x]$$

Equation (5.18) shows that $k(\nu)$ depends on the lineshape function $g(\nu)$; hence $I_\nu(\nu, x)$ is related exponentially to $g(\nu)$. Consequently the function $I_\nu(\nu, x)$ is much greater at the centre and smaller in the 'wings' than the atomic lineshape. $I_\nu(\nu, x)$ is therefore narrower than the atomic lineshape – this effect is known as spectral narrowing. In fact, as we shall see below, laser light has an even narrower spectral range than suggested by this argument.

5.8

Population inversion and pumping threshold conditions

We may now use eq. (5.18) to calculate the population inversion required to reach the lasing threshold. From eq. (5.18) we have

$$\left(N_2 - \frac{g_2}{g_1} N_1 \right) = \frac{k(\nu_s)c}{B_{21}h\nu_s n g(\nu_s)}$$

At threshold, the small signal gain coefficient is given by eq. (5.17), that is

$$k(\nu_s) = k_{th} = \gamma + \frac{1}{2L} \ln \left(\frac{1}{R_1 R_2} \right)$$

and therefore

$$\left(N_2 - \frac{g_2}{g_1} N_1 \right)_{th} = \frac{k_{th}c}{B_{21}h\nu_s n g(\nu_s)}$$

From eq. (5.8) we have $B_{21} = c^3 A_{21} / (8\pi h \nu^3 n^3)$. The quantity A_{21} may be determined experimentally by noting that it is the reciprocal of the spontaneous radiation lifetime τ_{21} from level 2 to level 1.

Thus combining the above equations we can write the threshold population inversion as

$$N_{th} = \left(N_2 - \frac{g_2}{g_1} N_1 \right)_{th} = \frac{8\pi \nu_{th}^2 k_{th} \tau_{21} n^2}{c^2 g(\nu_s)} \quad (5.22)$$

We note that the lasing threshold will be achieved most readily when $g(\nu_s)$ is a maximum, that is when ν_s has the value ν_0 corresponding to the centre of the natural linewidth. We may therefore replace $g(\nu_s)$ by $1/\Delta\nu$ (see eq. 5.21) to give

$$N_{th} = \frac{8\pi \nu_0^2 k_{th} \tau_{21} \Delta\nu n^2}{c^2} \quad (5.23)$$

We now proceed to calculate the pumping power required to reach threshold. To do this, we must solve the rate equations for the particular system. The rate equations describe the rate of change of the populations of the laser medium energy levels in terms of the emission and absorption processes and pump rate. We shall consider the ideal four-level system shown in Fig. 5.9. We assume that $E_1 \gg kT$ so that the thermal population of level 1 is negligible; we also assume that the threshold population density N_{th} is very small compared with the ground state population so that during lasing the latter is hardly affected. If we let \mathcal{R}_2 and \mathcal{R}_1 be the rates at which atoms are pumped into levels 2 and 1 respectively, we can write the rate equations for these levels (assuming $g_1 = g_2$ for simplicity, and hence $B_{21} = B_{12}$) as

$$\frac{dN_2}{dt} = \mathcal{R}_2 - N_2 A_{21} - \rho_\nu B_{21} (N_2 - N_1) \quad (5.24)$$

and

$$\frac{dN_1}{dt} = \mathcal{R}_1 + \rho_\nu B_{21} (N_2 - N_1) + N_2 A_{21} - N_1 A_{10} \quad (5.25)$$

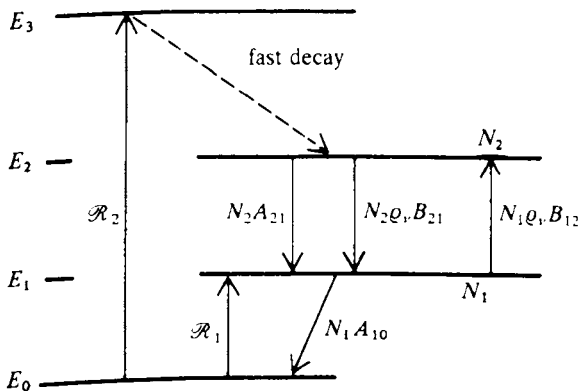


FIG. 5.9 Transitions within an ideal four-level system. (The principal pumping mechanism excites atoms to the level E_3 , from where they rapidly decay to the level of E_2 . As E_2 is the upper level of the laser transition, we have designated the corresponding pumping rate \mathcal{R}_2 rather than \mathcal{R}_1 .)

Process \mathcal{R}_1 , which populates the lower laser level 1, is detrimental to laser action as it clearly reduces the population inversion. Although such pumping is unavoidable in many lasers, for example gas lasers pumped via an electrical discharge (section 5.10.3), we shall henceforth ignore \mathcal{R}_1 . If we assume that the system is being pumped at a steady rate then we have $dN_2/dt = dN_1/dt = 0$. Hence we may solve eqs (5.24) and (5.25) for N_1 and N_2 . We leave it as an exercise for the reader to show that

$$N_1 = \mathcal{R}_2/A_{10}$$

$$N_2 = \mathcal{R}_2 \left(1 + \frac{\rho_v B_{21}}{A_{10}} \right) (A_{21} + \rho_v B_{21})^{-1}$$

and hence

$$N_2 - N_1 = \mathcal{R}_2 \left(\frac{1 - A_{21}/A_{10}}{A_{21} + \rho_v B_{21}} \right) \quad (5.26)$$

We can see that unless $A_{21} < A_{10}$, the numerator will be negative and no population inversion can take place. As the Einstein A coefficients are the reciprocals of the spontaneous lifetimes, the condition $A_{21} < A_{10}$ is equivalent to the condition $\tau_{10} < \tau_{21}$, that is the upper lasing level has a longer spontaneous emission lifetime than the lower level. In most lasers $\tau_{21} \gg \tau_{10}$ and $(1 - A_{21}/A_{10}) \approx 1$.

Now, below threshold we may neglect ρ_v since lasing has not yet commenced and most of the pump power appears as spontaneous emission; thus eq. (5.26) can be written as

$$N_2 - N_1 = \mathcal{R}_2 \left(\frac{1 - A_{21}/A_{10}}{A_{21}} \right)$$

That is, there is a linear increase in population inversion with pumping rate but insufficient inversion to give amplification.

At threshold, ρ_v is still small and assuming $g_1 = g_2$ we can express the threshold population inversion in terms of the threshold pump rate, that is

$$(N_2 - N_1)_{\text{th}} = N_{\text{th}} = \mathcal{R}_{\text{th}} \left(\frac{1 - A_{21}/A_{10}}{A_{21}} \right) \quad (5.27)$$

or inserting the above approximation that $(1 - A_{21}/A_{10}) \approx 1$

$$\mathcal{R}_{\text{th}} = N_{\text{th}} A_{21}$$

or

$$\mathcal{R}_{\text{th}} = N_{\text{th}}/\tau_{21}$$

Each atom raised into level 2 requires an amount of energy E_2 so that the total pumping power per unit volume P_{th} required at threshold may be written as

$$P_{\text{th}} = E_2 N_{\text{th}}/\tau_{21}$$

We may substitute for N_{th} from eq. (5.23) to give

$$P_{\text{th}} = \frac{E_s 8\pi v_0^2 k_{\text{th}} \tau_{21} \Delta v n^2}{\tau_{21} c^2}$$

or

$$P_{\text{th}} = \frac{E_s 8\pi v_0^2 k_{\text{th}} \Delta v n^2}{c^2} \quad (5.28)$$

This is the point at which the gain due to the population inversion exactly equals the cavity losses. Further increase of the population inversion with pumping is impossible in a steady state situation, since this would result in a rate of induced energy emission which exceeds the losses. Thus the total energy stored in the cavity would increase with time in violation of the steady state assumption (this is the phenomenon of gain saturation described earlier).

This argument suggests that $[N_2 - (g_2/g_1)N_1]$ must remain equal to N_{th} regardless of the amount by which the threshold pump rate is exceeded. Equation (5.26) shows that this is possible providing $\rho_v B_{21}$ is able to increase (once \mathcal{R}_2 exceeds its threshold value given by eq. 5.27) so that the equality

$$N_{\text{th}} = \mathcal{R}_2 \left(\frac{1 - A_{21}/A_{10}}{A_{21} + \rho_v B_{21}} \right)$$

is satisfied. Now combining this equation with eq. (5.27) we have

$$\frac{\mathcal{R}_{\text{th}}}{A_{21}} = \frac{\mathcal{R}_2}{A_{21} + \rho_v B_{21}}$$

and hence

$$\rho_v = \frac{A_{21}}{B_{21}} \left(\frac{\mathcal{R}_2}{\mathcal{R}_{\text{th}}} - 1 \right) \quad (5.29)$$

Since the power output W of the laser will be directly proportional to the optical power density within the laser cavity and the pump rate into level 2 (i.e. \mathcal{R}_2) will be proportional to the pump power P delivered to the laser, we may rewrite eq. (5.29) as

$$W = W_0 \left(\frac{P}{P_{\text{th}}} - 1 \right) \quad (5.30)$$

where W_0 is a constant.

Thus if the pump rate is increased above the value P_{th} the beam irradiance is expected to increase linearly with pump rate. This is borne out in practice and plots of population inversion and laser output as a function of pump rate are of the form shown in Fig. 5.10. The additional power above threshold is channelled into a single (or a few) cavity mode(s) (see section 5.9). Spontaneous emission still appears above threshold but it is extremely weak in relation to the laser output as it is emitted in all directions and has a much greater frequency spread.

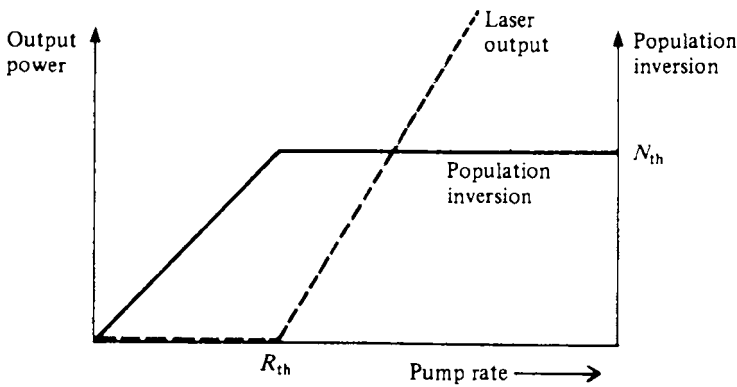


FIG. 5.10 Population inversion and laser power output as a function of pump rate.

5.9

Laser modes

Examination of a laser output with a spectrometer of very high resolving power, such as the scanning Fabry–Perot interferometer, reveals that it consists of a number of discrete frequency components (or very narrow spectral lines). To appreciate how these discrete lines arise and how they are related to the laser transition lineshape we need to examine the effects of the mirrors on the light within the laser cavity (see refs 5.8 and 5.2a).

5.9.1 Axial modes

The two mirrors of the laser form a resonant cavity and standing wave patterns are set up between the mirrors in exactly the same way that standing waves develop on a string or within an organ pipe. The standing waves satisfy the condition

$$p \frac{\lambda}{2} = L$$

or

$$\nu = \frac{pc}{2L} \quad (5.31)$$

where, strictly speaking, L is the *optical* path length between the mirrors, in which case λ_0 would be the vacuum wavelength and p is an integer, which may be very large (e.g. if $L = 0.5$ m and $\lambda \approx 500$ nm then $p \approx 2 \times 10^6$). As p has such a large value, many different values of p are possible for only a small change in wavelength. Each value of p satisfying eq. (5.31) defines an *axial* (or longitudinal) *mode* of the cavity.

From eq. (5.31), the frequency separation $\delta\nu$ between adjacent modes ($\delta p = 1$) is given by

$$\delta\nu = \frac{c}{2L} \quad (5.32)$$

and therefore for $L = 0.5$ m, $\delta\nu = 300$ MHz. As eq. (5.32) is independent of p , the frequency separation of adjacent modes is the same irrespective of their actual frequencies. The modes of oscillation of the laser cavity will consist, therefore, of a large number of frequencies, each given by eq. (5.31) and separated by $c/2L$, as illustrated in Fig. 5.11(b).

It should be appreciated, however, that while all the integers p give *possible* axial cavity

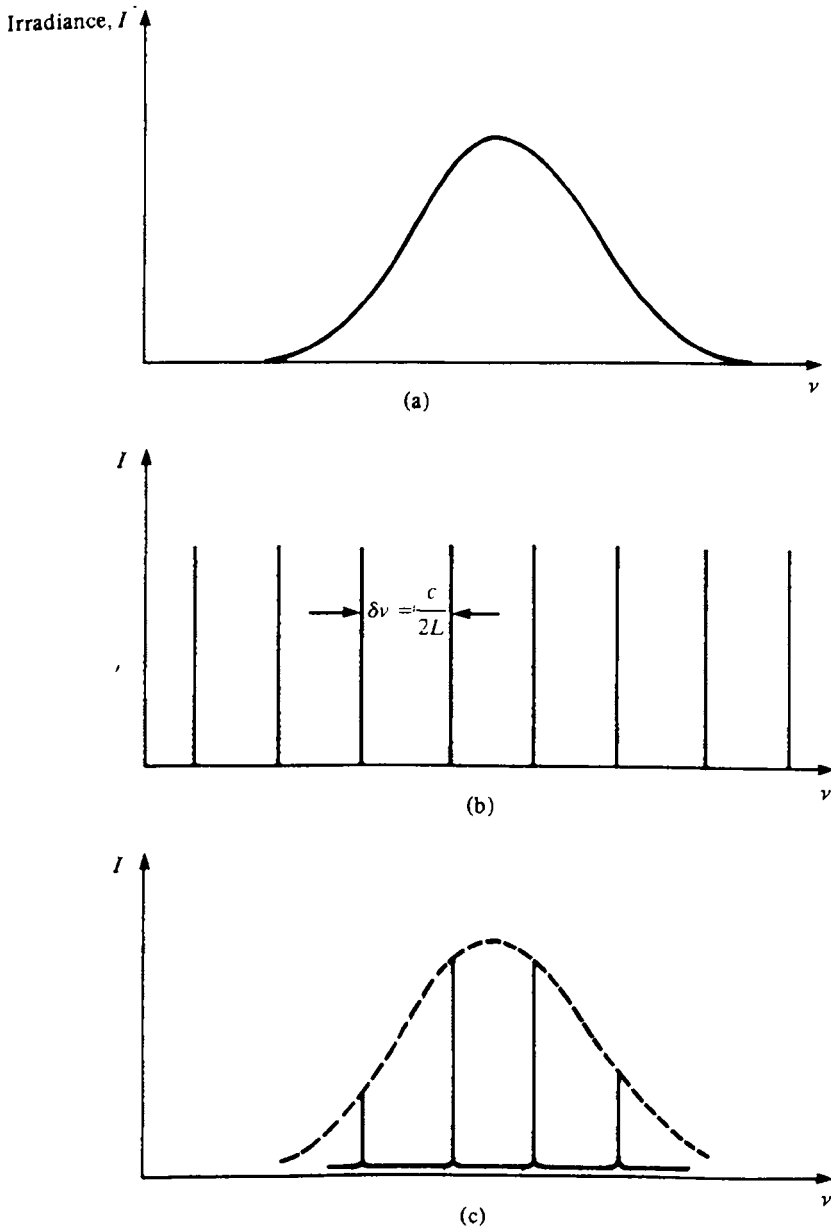


FIG. 5.11 Broadened laser transition line (a), cavity modes (b), and axial modes in the laser output (c).

modes only those which lie within the gain curve or laser transition line will actually oscillate. Thus the broadened laser transition (Fig. 5.11a) for the 632.8 nm wavelength emitted by neon is about 1.5×10^9 GHz wide so that with the 0.5 m long cavity in the above example we would expect four or five modes to be present as illustrated in Fig. 5.11(c). Figure 5.12 shows the axial modes of an He–Ne laser about 1 m in length which are displayed with the aid of an optical frequency analyzer and silicon photodiode.

The axial modes all contribute to a single ‘spot’ of light in the laser output, whereas the transverse modes discussed below may give rise to a pattern of spots in the output. If the linewidth of the axial modes is measured it will be found to be much narrower than the width of the Fabry–Perot resonances to be expected from treating the cavity simply as a Fabry–Perot interferometer (see ref. 5.9). We can appreciate the reason for this by considering the quality factor Q of the resonator. Q can be defined in general by

$$Q = \frac{2\pi \times \text{energy stored in the resonator}}{\text{energy dissipated per cycle}}$$

or

$$Q = \frac{\text{resonant frequency}}{\text{linewidth}} = \frac{\nu}{\Delta\nu}$$

For an electrical oscillator, Q may be approximately 100. However, for a laser, Q may be about 10^8 and hence $\Delta\nu \approx 1$ MHz, which is much narrower than the Fabry–Perot resonances, which are about 10^9 Hz. Indeed, in lasers the active medium is actually supplying energy

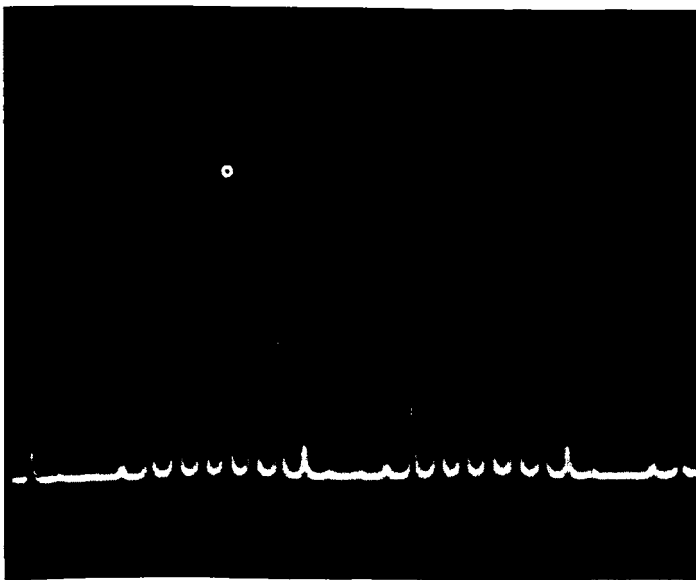


FIG. 5.12 Axial modes formed in the He–Ne laser – the mode pattern is repeated (here, three times) as the optical frequency analyzer scans through the gain curve of the laser. (Photograph courtesy of Dr I. D. Latimer, School of Engineering, University of Northumbria at Newcastle.)

to the oscillating modes so that in theory the energy dissipation can be zero and Q infinite. In practice, there are always losses which prevent this happening, but even so linewidths of about 1 Hz have been achieved.

5.9.2 Transverse modes

Axial modes are formed by plane waves travelling axially along the laser cavity on a line joining the centres of the mirrors. For any real laser cavity there will probably be waves travelling just off axis that are able to replicate themselves after covering a closed path such as that shown in Fig. 5.13(a). These will also give rise to resonant modes, but because they have components of their electromagnetic fields which are transverse to the direction of propagation they are termed transverse electromagnetic (or TEM) modes. A complete analysis of TEM modes is quite complicated and will not be attempted here (but see Chapter 8 for a further discussion). They are characterized by two integers q and r so that, as Fig. 5.13(b) shows, we have TEM_{00} , TEM_{01} , TEM_{11} , etc., modes (q gives the number of minima as the beam is scanned horizontally and r the number of minima as it is scanned vertically).

In a TEM_{00} mode, the irradiance distribution across the beam is in fact Gaussian, and so we may write the electric field variation as

$$\mathcal{E}(x, y) = \mathcal{E}_0 \exp\left(-\frac{x^2 + y^2}{w^2}\right) \quad (5.33)$$

where x and y are measured in directions perpendicular to the laser axis which is taken to be along the z direction. The sideways spread of the beam is determined by the value of the parameter w , which is a function of the distance z . When $x^2 + y^2 > w^2$, the field falls off rapidly with distance away from the laser axis. The value of w is determined by the locus of points where the field amplitude has fallen to $\exp(-1)$ of its maximum value (i.e. where $x^2 + y^2 = w^2$). Figure 5.14 shows the typical variation of w , with position, within a cavity formed by two concave mirrors of radius of curvature r_1 and r_2 separated by L . Such a cavity can be shown to be stable when $L \leq r_1 + r_2$. The surfaces of constant phase are not in general plane, but are perpendicular to the contours of constant field strength. It can be seen that the mirrors themselves are surfaces of constant phase. This is no accident, but merely a direct consequence of the requirement that the mode be self-replicating as the light energy flows backwards and forwards between the mirrors. At one position within the cavity, the wavefronts become plane and, in fact, a plane mirror placed at this point would give rise to a hemispherical cavity. At this point also w has its smallest value, w_0 . The variation of w with z is given by (ref. 5.10)

$$w(z) = w_0 \left[1 + \left(\frac{z\lambda}{\pi w_0^2} \right)^2 \right]^{1/2} \quad (5.34)$$

where z is measured from the position of minimum beam diameter. The precise value of w_0 depends on the type of cavity. In the case of a nearly confocal cavity where $r_1 = r_2 = r \geq L$ it can be shown that (ref. 5.5)

$$w_0^2 \approx \frac{\lambda}{2\pi} [L(2r - L)]^{1/2} \quad \text{or} \quad w_0^2 \approx \frac{\lambda r}{2\pi} \quad (5.35)$$

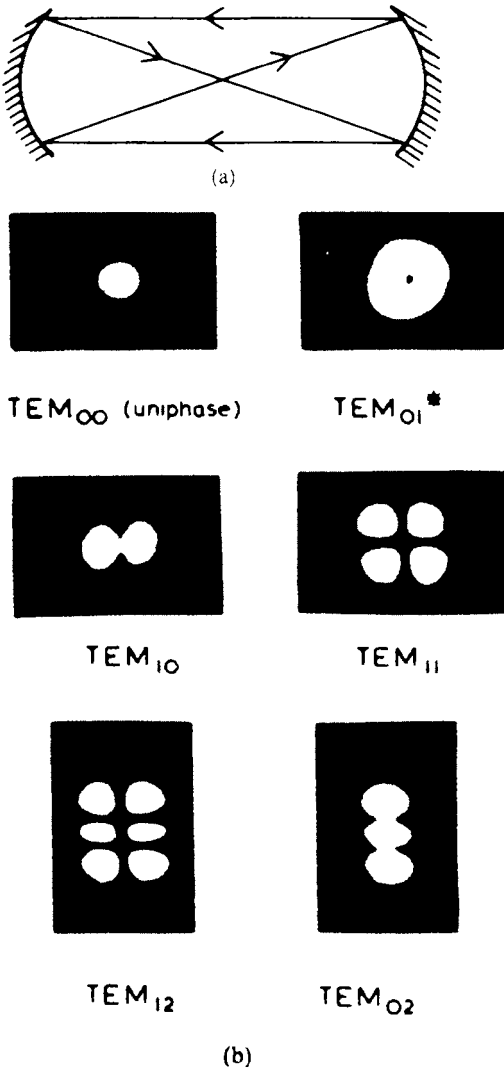


FIG. 5.13 (a) Example of a non-axial self-replicating ray that gives rise to transverse modes. (b) Some low order transverse modes of a laser. The modes are designated TEM_{qr}, where q and r are integers referring to the number of minima as the laser beam is scanned horizontally and vertically. The TEM₀₁^{*} mode is a combination of TEM₀₁ and TEM₁₀ modes. (From M. J. Beesley, *Lasers and their Applications*, 1972; courtesy of Taylor and Francis Ltd.)

Combining eqs (5.34) and (5.35) yields

$$w(z) = w_0 \left[1 + \left(\frac{2z}{r} \right)^2 \right]^{1/2} \tag{5.36}$$

Equation (5.36) applies outside as well as inside the cavity, and at large distances from the

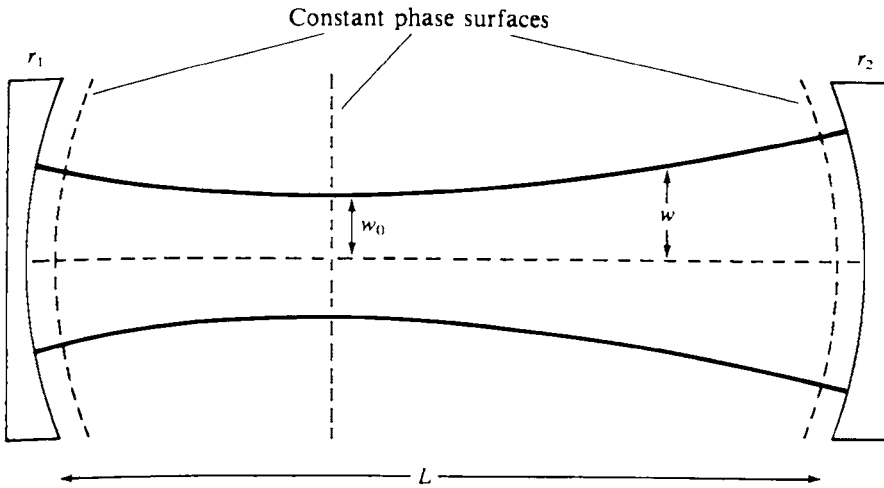


FIG. 5.14 TEM_{00} mode within a laser cavity. The mode 'adjusts' itself so that the mirror surfaces are surfaces of constant phase. The value of w is determined by the locus of points where the field amplitude has fallen to $\exp(-1)$ of its maximum value.

laser, that is when $z \gg r$, we have

$$w(z) = \frac{z\lambda}{\pi w_0} \quad (5.37)$$

As the TEM_{01} and higher order modes extend further from the cavity axis than the TEM_{00} mode, they will only oscillate if the aperture of the cavity is large enough. These and higher order modes can, therefore, be eliminated by narrowing the laser cavity leaving just the TEM_{00} mode oscillating. The TEM_{00} mode is often called the uniphase mode, since all parts of the propagating wavefront are in phase; this is not the case with higher order modes and in fact phase reversals account for the higher order transverse mode patterns. Consequently, a laser operating only in the TEM_{00} mode has the greatest spectral purity and degree of coherence, while operation in multimode form provides considerably more power.

It should be noted that each transverse mode will have the axial modes discussed above associated with it, so that the total spread in the laser spectrum may be (relatively) large.

5.10 Classes of laser

In the 40 years since Maiman reported the first observation of successful laser action in ruby, there has been an extremely rapid increase in the types of lasers and in the range of materials in which lasing has been shown to occur. It is not possible to describe all of these developments, so in this section we have concentrated on a description of the construction and mode of operation of some of the more commonly available and important lasers. These are classified into four groups: doped insulator, semiconductor, gas and dye lasers.

Before discussing these laser types it might be useful to remind ourselves of the basic requirements which must be satisfied for laser operation.

First, there must be an active medium which emits radiation in the required region of the electromagnetic spectrum. Secondly, a population inversion must be created within the medium; this, in turn, requires the existence of suitable energy levels associated with the lasing transition for pumping. Thirdly, for true laser oscillation there must be optical feedback at the ends of the medium to form a resonant cavity (satisfying the first two conditions can provide light amplification but not the highly collimated, monochromatic beam of light which makes lasers so useful).

5.10.1 Doped insulator lasers

The term doped insulator laser is used to describe a laser whose active medium consists of a crystalline or amorphous (glassy) host material containing *active* ions, typically from the transition metal and rare earth groups of elements in the periodic table. These ions are impurities which are intentionally introduced (i.e. doped) either into the crystal during its growth, or to the melt from which the glass solidifies. In addition to the wide range of ions mentioned above there are several suitable host materials available, such as sapphire (Al_2O_3), garnets, aluminates and fluorides, thereby providing a large range of lasing wavelengths, together with associated pumping transitions.

The principal characteristics of suitable host crystals, normally in the form of a rod, rectangular bar or fiber include:

- high degree of optical uniformity and freedom from defects which act as scattering centres;
- low expansivity and high thermal conductivity to reduce thermal stress, and to maintain a uniform refractive index when heated during the pumping process;
- easy crystal growth with lattice sites which readily accept the impurity ions at high doping levels, up to approximately 10^{26} m^{-3} .

Glasses are much easier to fabricate with uniform composition and optical properties, and can be easily made into a variety of different shapes. They can also normally be doped to an even higher concentration than the crystalline hosts.

Doped insulator lasers are rugged, easy to maintain and capable of generating high peak powers. They are invariably optically pumped and the availability of suitable pumping sources of light is an important consideration. Typical examples are the ruby, Nd:YAG, alexandrite, YLF and silicate glass lasers. Although the ruby laser is interesting in that it was the first successful laser, the Nd:YAG laser is now much more widely used and we describe it in some detail, giving only passing reference to the ruby and other lasers.

Nd:YAG LASER

The active medium for this laser is yttrium aluminium garnet ($\text{Y}_3\text{Al}_5\text{O}_{12}$) with the rare earth metal ion neodymium Nd^{3+} present as an impurity. The Nd^{3+} ions, which are randomly distributed as substitutional impurities on lattice sites normally occupied by the yttrium ions, provide the energy levels for both the lasing transitions and pumping. Though the YAG host itself does not participate directly in the lasing action, it does have two important roles.

When an Nd^{3+} ion is placed in a host crystal lattice it is subject to the electrostatic field of the surrounding ions, the so-called *crystal field*. The crystal field of the host interacts with the electron energy levels in a variety of ways depending on such factors as its strength and symmetry, and on the electron configuration of the impurity.

A neodymium ion which is free to move in a gaseous discharge, for example, has many of its energy levels with the same energy; these are said to be degenerate. When the ion is placed in the host the crystal field splits some of the energy levels, thereby partly removing the degeneracy. A rather more important effect in the case of the Nd:YAG laser is that the crystal field modifies the transition probabilities between the various energy levels of the Nd^{3+} ion so that some transitions, which are forbidden in the free ion, become allowed.

The net result is that the ground and first excited state energy levels of the Nd^{3+} ion split into the groups of levels shown in Fig. 5.15. The symbols used to describe the energy levels in this and succeeding diagrams depend on the exact nature of the ions and atoms involved. In the case under discussion, the symbols arise from Russell–Saunders or LS coupling (ref. 5.11). The symbol for an energy level is written $^{2S+1}X_J$. Here S is the vector sum of the electron spins of the ion. X gives the vector sum L of the orbital angular momentum quantum numbers, where values of $L = 0, 1, 2, 3, 4, \dots$ are designated by S, P, D, F, G, Finally J is

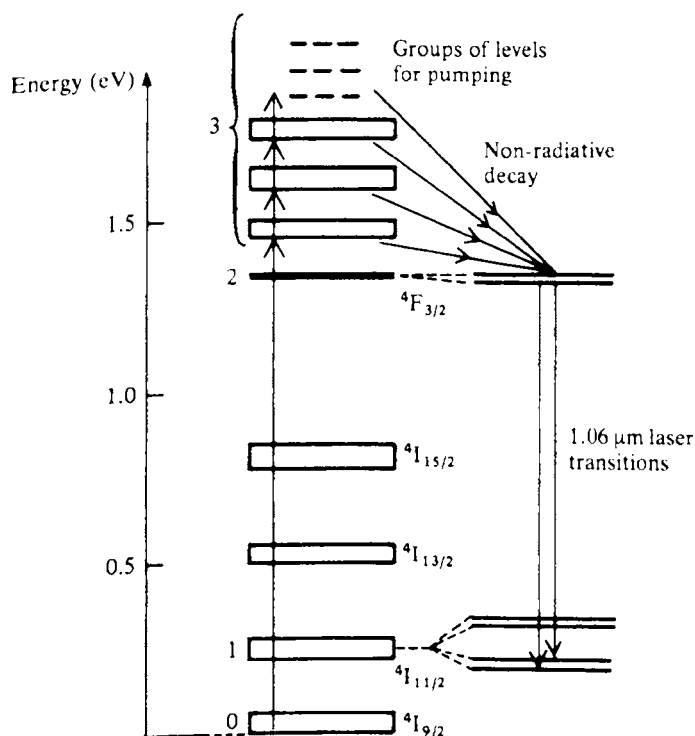


FIG. 5.15 Simplified energy level diagram² for the neodymium ion in YAG showing the principal laser transitions. Laser emission also results from transitions between the $^4F_{3/2}$ levels and the $^4I_{13/2}$ and $^4I_{15/2}$ levels, but at only one-tenth of the intensity of the transition shown.

the vector sum of S and L (see ref. 5.11). Thus in the ${}^4F_{3/2}$ level, $S = 3/2$ (there are three unpaired electrons in the $4f$ subshell of the Nd^{3+} ion), $L = 3$ and $J = 3/2$. Whilst familiarity with this and similar notations for electron energy levels is not essential to appreciate the basics of a given laser, it is a prerequisite for a detailed understanding of the mechanisms involved.

As we can see from Fig. 5.15 the Nd:YAG laser is essentially a four-level system, that is the terminal laser level ${}^4I_{11/2}$ is sufficiently far removed from the ground state ${}^4I_{9/2}$ that its room temperature population is very small. Whilst a number of laser transitions may occur between some of the pairs of levels shown to the right of the figure, the most intense line at $1.064\text{ }\mu\text{m}$ arises from the superposition of the two transitions shown.

Pumping is normally achieved by using an intense flash of white light from a xenon flashtube. This excites the Nd^{3+} ions from the ground state to the various energy states above the ${}^4F_{3/2}$ state; there are, in fact, many more states at higher energies than are shown in Fig. 5.15. The presence of several possible pumping transitions contributes to the efficiency of the laser when using a pumping source with a broad spectral output. To ensure that as much radiation as possible from the flashtube is absorbed in the laser medium, close optical coupling is required. The usual arrangement is shown in Fig. 5.16; a linear flashtube and the lasing medium in the form of a rod are placed inside a highly reflecting ellipsoidal cavity. If the flashtube is along one focal axis and the laser rod along the other, then the properties of the ellipse ensure that most of the radiation from the flashtube passes through the laser. The flashtube is fired by discharging a capacitor bank through the tube; the discharge is often initiated by using a secondary high voltage ($\approx 20\text{ kV}$) trigger pulse.

As the pumping flash lasts for only a short time ($\approx 1\text{ ms}$) the laser output is in the form of a pulse, which starts about 0.5 ms after the pumping flash starts. This represents the time for the population inversion to build up. Once started, stimulated emission builds up rapidly

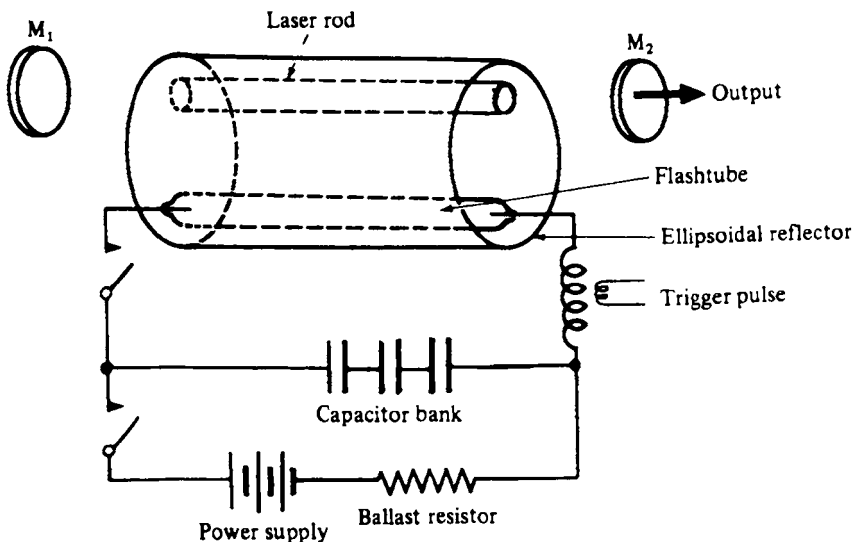


FIG. 5.16 Typical construction of a doped insulator laser showing the ellipsoidal reflector used to maximize optical coupling between the flashtube and laser rod.

and thus depopulates the upper lasing level 2 – much faster in fact than the pumping can replace the excited atoms so that laser action momentarily stops until population inversion is achieved again. This process then repeats itself so that the output consists of a large number of spikes of about $1\ \mu\text{s}$ duration with about $1\ \mu\text{s}$ separation. As the spikes are unrelated, the coherence of the laser pulse (which lasts a total of about $0.5\ \text{ms}$), while being much greater than that of classical light sources, is not as great as might be expected.

The optical cavity may be formed by grinding the ends of the Nd:YAG rod flat and parallel and then silvering them. More usually, however, external mirrors are used as shown in Fig. 5.16. One mirror is made totally reflecting while the other is about 10% transmitting to give an output.

A large amount of heat is dissipated by the flashtube and consequently the laser rod quickly becomes very hot. To avoid damage resulting from this, and to allow a reasonable pulse repetition rate, cooling has to be provided by forcing air over the crystal using the reflector as a container. For higher power lasers it is necessary to use water cooling. Provided sufficient cooling is available it is possible to replace the xenon flashtube by a krypton one or a quartz–halogen lamp and to operate the laser continuously.

A glance at the energy level diagram, Fig. 5.15, shows that the maximum possible power efficiency of the laser, ν_{21}/ν_{03} , is about 80%. In practice, because of losses in the system (which include the loss in converting electrical to optical energy in the pumping source, the poor coupling of the pumping source output to the laser rod and the small fraction of the pumping radiation which is actually absorbed), the actual power efficiency is typically 0.1%. Thus, a laser pumped by a flash lamp operated by the discharge of a $1000\ \mu\text{F}$ capacitor charged to 4–5 kV (i.e. an input energy of about 10 kJ) may produce an output pulse of about 10 J. As the pulse lasts for only about $0.5\ \text{ms}$, the average power, however, is then about $2 \times 10^4\ \text{W}$. The peak power may be greatly increased by the technique of *Q*-switching, which is described in section 6.4.

Nd:GLASS LASER

Both silicate and phosphate glasses have high optical homogeneity and provide excellent host materials for neodymium. Local electric fields within the glass modify the Nd^{3+} ion energy levels in much the same way as the crystal field in YAG. The Nd^{3+} ions, however, may be in a variety of slightly different environments causing the linewidth to be much broader than in YAG by a factor of about 50, thereby raising the threshold pumping power required for laser action (eq. 5.28 showed that the threshold pump power is proportional to $\Delta\nu$). In consequence, Nd:glass lasers are operated in the pulsed mode and the output spectral linewidth is greater than in Nd:YAG. On the other hand, glass can be doped more heavily than YAG (6% as opposed to 1.5%) and up to three times as much energy can therefore be produced by Nd:glass lasers; it is also much easier and cheaper to prepare glass rods than to grow YAG crystals.

RUBY LASER

The basic principle of operation of the ruby laser is the same as that of the Nd:YAG laser. The active medium is a synthetically grown crystal of ruby, that is aluminium oxide, with about 0.05% by weight of chromium as an impurity. Chromium ions, Cr^{3+} , replace aluminium ions

in the lattice and the crystal field partially removes the degeneracy of the isolated ions to provide levels for pumping and for the laser transitions. In this case, some of the energy levels of the Cr^{3+} ions are almost independent of the crystal field and they remain sharp. Other levels, however, are strongly dependent on the crystal field so that lattice vibrations, which cause fluctuations in the crystal field, broaden these levels quite substantially. The ${}^2\text{E}$ and ${}^4\text{A}_2$ levels remain sharp while the ${}^4\text{T}_1$ and ${}^4\text{T}_2$ levels become broad as shown in Fig. 5.17 (the symbols in this case are derived from crystal field theory, ref. 5.12). Thus the pump transitions are spectrally broad while the transitions R_1 and R_2 are narrow. The energy level diagram shows that ruby is basically a three-level system. As explained previously, rather more than half of the total number of ions have to be pumped to level 2 via level 3 to create a population inversion. Thus the laser has a very low efficiency compared with a four-level system such as $\text{Nd}:\text{YAG}$. Pumping is achieved through the absorption of the green and blue spectral regions of a white light discharge; this absorption, of course, accounts for the colour of ruby.

DIODE-PUMPED LASERS

It was mentioned earlier in this section that, because of poor optical coupling and a poor match of the pumping source emission spectrum to the absorption bands of the Nd^{3+} ion,

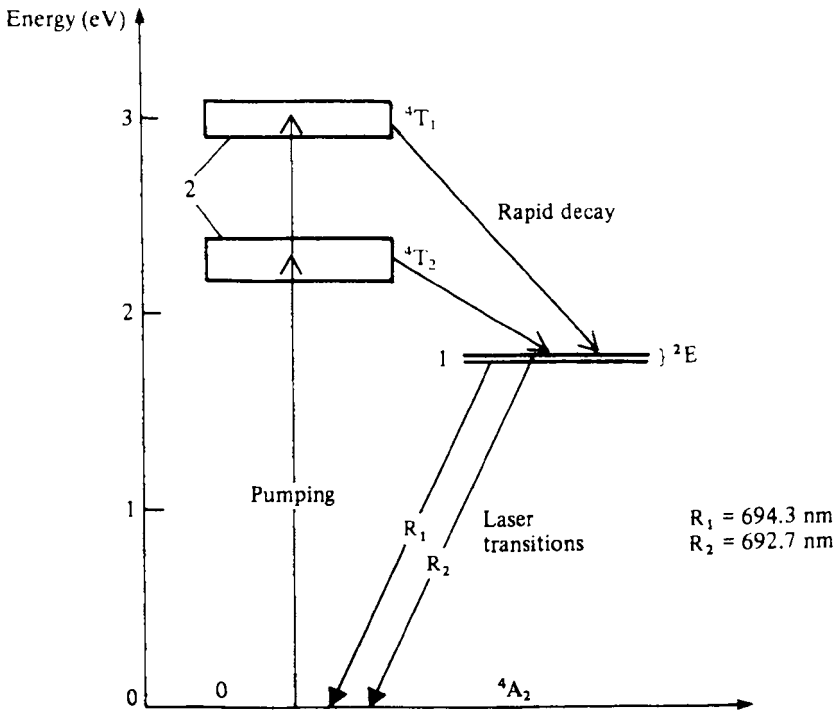


FIG. 5.17 Three-level ruby laser system. Pumping is due to the Cr^{3+} ions absorbing blue (excitation to ${}^4\text{T}_1$ levels) and green light (excitation to ${}^4\text{T}_2$ levels). The wavelengths of the R_1 and R_2 laser lines are temperature dependent, the values given are typical. (Some energy levels not directly involved in the laser transition have been omitted for clarity.)

the overall efficiency of Nd:YAG and glass lasers, despite having four-level energy level systems, is less than 1%. Significant improvement of the overall efficiency can be achieved by pumping the Nd:YAG with semiconductor diode lasers, which as we shall see in section 5.10.2 are themselves very efficient. Although we have not yet discussed such lasers, for the present purpose we can regard them as sources of nearly monochromatic radiation whose emission wavelength can be tuned by, for example, changing their temperature. GaAs/GaAlAs lasers can emit at a wavelength of 808 nm which coincides with a very strong absorption peak in Nd:YAG, so that very efficient pumping can be achieved. This readily permits CW operation without the need for water cooling, with the added benefit of a substantial reduction in the overall size of the laser and much improved frequency stability (see section 6.1).

Pumping is usually carried out longitudinally, that is the light from the laser diodes enters through one or both ends of the YAG crystal as shown in Fig. 5.18. The cavity end mirrors have very high reflectances at the laser wavelength of $1.06\ \mu\text{m}$, but are transparent at the pumping wavelength of 808 nm. Even more compact devices can be made in which dielectric reflecting coatings are applied to the ends of the laser rod.

With the advent in recent years of high power laser diode bars and arrays (see section 5.10.2) comprising several devices in a linear or two-dimensional array operating simultaneously, CW outputs of 20 W or more can be achieved from diode-pumped Nd:YAG lasers. In these cases the light from the diode arrays can be efficiently collected using optical fibers which transform the elliptical cross-section of the output beam of laser diodes into a circular beam which is suitable for end pumping. In this way some 65% of the output of the diode array can be coupled into the laser crystal. Alternatively the laser crystal can be pumped transversely using many diode arrays arranged symmetrically around the laser rod (ref. 5.13).

5.10.1.1 Glass fiber laser

An interesting extension of the idea of an end-pumped solid rod laser is the fiber laser (ref. 5.14). These use optical fiber instead of solid rods. Optical fibers will be dealt with in detail in Chapter 8; however, here we may simply regard them as being small diameter 'flexible' rods consisting of a narrow diameter 'core' region surrounded by a somewhat thicker 'cladding'. Light may readily propagate down the core region of such fibers with minimal sideways loss, even when the fiber is bent. The core of the fibers, which are usually made from silica (SiO_2), may be doped with Nd^{3+} to create a lasing medium. A laser cavity may

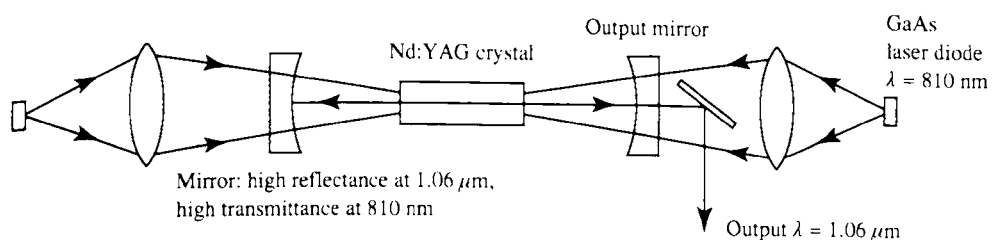


FIG. 5.18 Diagram of a longitudinal or end-pumped, diode-pumped, solid state laser – here Nd:YAG.

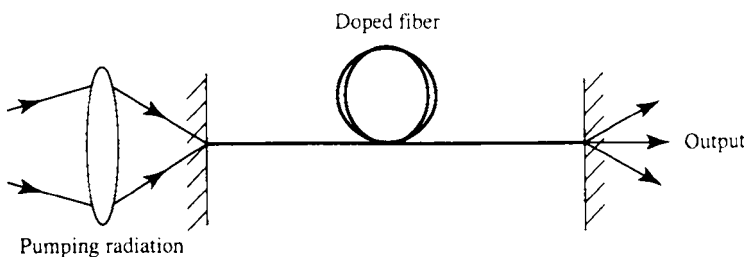


FIG. 5.19 End-pumped fiber laser.

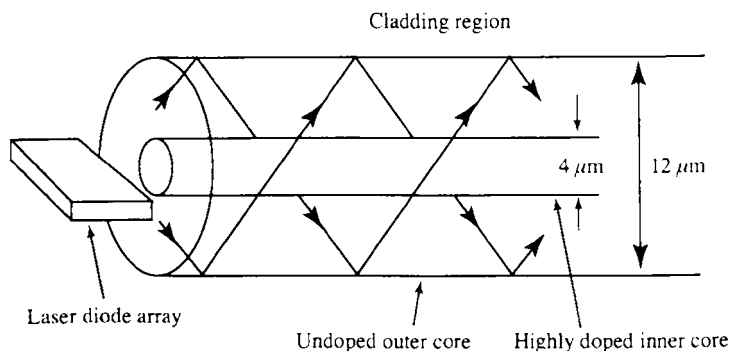


FIG. 5.20 Cladding-pumped fiber laser – the pumping radiation is effectively captured by the outer core and coupled into the inner core.

be formed by butting dielectric mirrors up against the ends of the fiber as shown in Fig. 5.19. The so-called cladding pumping technique (Fig. 5.20) helps to overcome the difficulty of efficiently coupling the laser light into the extremely narrow fiber core, by using an outer, undoped, core of much larger area to collect the pump light. Nd-doped fiber lasers of this type may emit 5 W or more in a single transverse mode. It should perhaps be noted that in relation to the fibers described in Chapter 8 cladding pumping is a misnomer as in reality the light is collected by an outer core and not the cladding. Alternatively the pumping light, at a wavelength of 810 nm, may be introduced into the fiber core using one of the techniques described in section 8.5. While there are several applications for fiber lasers their most significant contribution to date has been in fiber optic communications, where fiber amplifiers, rather than oscillators, are beginning to make a major contribution (see section 9.3).

5.10.1.2 Vibronic lasers

Vibronic lasers such as alexandrite and Ti:sapphire are superficially similar to other solid state lasers such as Nd:YAG in that light from an external pump source excites impurity ions in a transparent host. They are fundamentally different, however, in that laser gain is possible over a broad range of wavelengths so that they can produce either tunable outputs or ultrashort pulses (see Chapter 6). The reason for this is that the electronic energy levels

involved in the laser transition in vibronic lasers are broadened into bands of vibrational sub-levels, corresponding to discrete quantized lattice vibration energies. Thus transitions can take place over a range of energy differences between the upper and lower levels giving a corresponding range of wavelengths. This behaviour is rather like that of the dye lasers discussed in section 5.10.4.

Most vibronic solid state lasers have four-level systems that operate in much the same way (ref. 5.15). The pumping radiation excites the active ions to a vibronic band. The ions then lose vibrational energy and drop to the bottom of the band, which is the upper laser level. The laser transition then occurs to a vibrationally excited sublevel of the ground electronic state; this is followed by the ions relaxing to the lowest sublevel of the ground state by releasing vibrational energy as illustrated in Fig. 5.21, which also shows a fixed wavelength transition.

Many vibronic lasers use chromium ions as the active ingredient and it is instructive to contrast the ruby and alexandrite laser which comprises Cr^{3+} ions in a beryllium aluminate (BeAl_2O_4) host. In the case of ruby, as we saw earlier, the energy levels are discrete and we have a three-level system. In contrast Fig. 5.21 shows that the ${}^4\text{T}_2$ and ${}^4\text{A}_2$ levels are broadened and that in effect we have a four-level system. Pumping at wavelengths between 380 nm

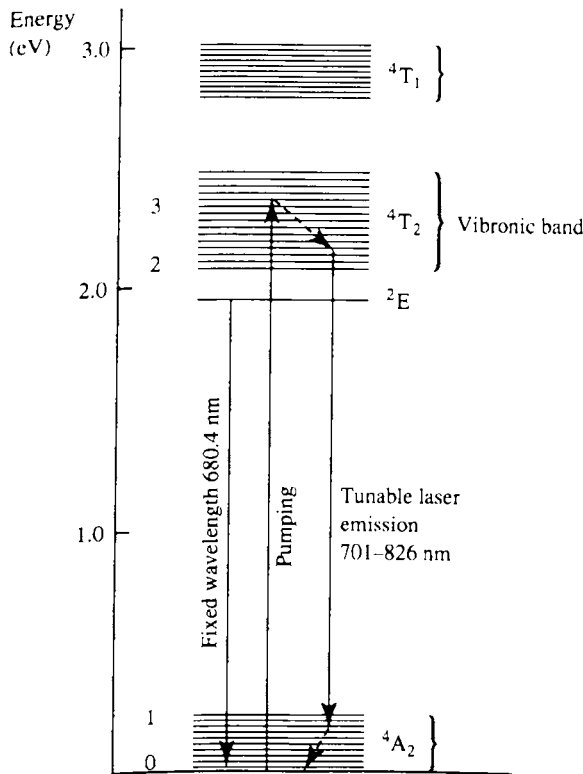


FIG. 5.21 Energy level diagram for Cr^{3+} ions in alexandrite (BeAl_2O_4) – some energy levels not directly involved in the laser transitions have been omitted for clarity.

and 630 nm excites the Cr^{3+} ions. Laser action then occurs between the bottom of the ${}^4\text{T}_2$ band to one of the sublevels in the ${}^4\text{A}_2$ band giving rise to emission wavelengths in the range 701–826 nm.

In the absence of any wavelength-selective element in the optical system the laser should oscillate at the peak of the gain curve. Tuning at a particular wavelength is often accomplished by incorporating a birefringent filter within the optical cavity: the filter is rotated to pass a very narrow range of wavelengths within the tuning range.

5.10.2 Semiconductor lasers

Semiconductor lasers are not very different in principle from the light-emitting diodes discussed in Chapter 4. A p–n junction provides the active medium: thus, to obtain laser action we need only meet the other necessary requirements of population inversion and optical feedback. To obtain stimulated emission, there must be a region of the device where there are many excited electrons and vacant states (i.e. holes) present together. This is achieved by forward biasing a junction formed from very heavily doped n and p materials. In such n^+ -type material, the Fermi level lies within the conduction band. Similarly, for the p^+ -type material the Fermi level lies in the valence band. The equilibrium and forward-biased energy band diagrams for a junction formed from such so-called degenerate materials are shown in Fig. 5.22. When the junction is forward biased with a voltage that is nearly equal to the energy

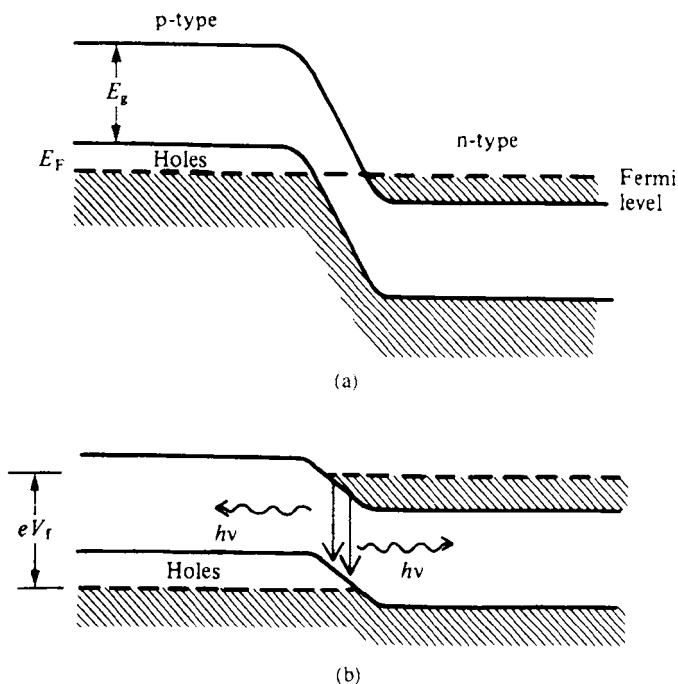


FIG. 5.22 Heavily doped p–n junction: (a) in equilibrium and (b) with forward bias (the dashed lines represent the Fermi level in equilibrium (a) and with forward bias (b)).

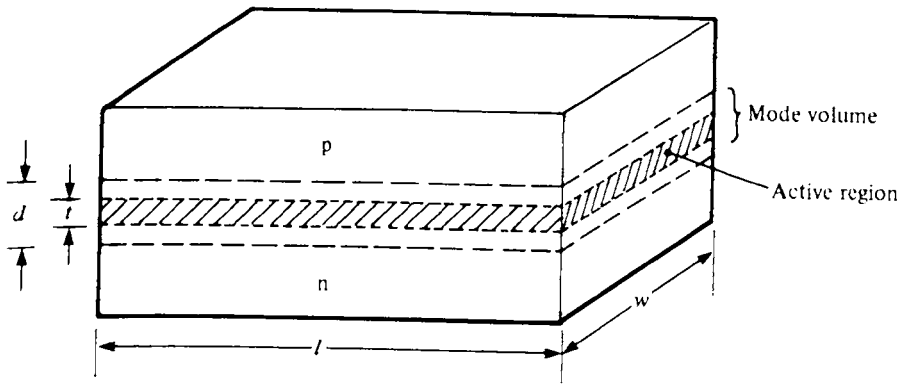


FIG. 5.23 Diagram showing the active region and mode volume of a semiconducting laser.

gap voltage E_g/e , electrons and holes are injected across the junction in sufficient numbers to create a population inversion in a narrow zone called the *active region* (Fig. 5.23).

The thickness t of the active region can be approximated by the diffusion length L_e of the electrons injected into the p region, assuming that the doping level of the p region is less than that of the n region so that the junction current is carried substantially by electrons (sections 2.7.1 and 2.8.2). For heavily doped GaAs at room temperature L_e is 1–3 μm .

In the case of those materials such as GaAs which have a direct bandgap (section 2.4) the electrons and holes have a high probability of recombining radiatively. The recombination radiation produced may interact with valence electrons and be absorbed, or interact with electrons in the conduction band thereby stimulating the production of further photons of the same frequency ($\nu = E_g/h$). If the injected carrier concentration becomes large enough, the stimulated emission can exceed the absorption so that optical gain can be achieved in the active region. Laser oscillations occur, as usual, when the round trip gain exceeds the total losses over the same distance. In semiconductors, the principal losses are due to scattering at optical inhomogeneities in the semiconductor material and free carrier absorption. The latter results when electrons and holes absorb a photon and move to higher energy states in the conduction band or valence band respectively. The carriers then return to lower energy states by non-radiative processes.

In the case of diode lasers, it is not necessary to use external mirrors to provide positive feedback. The high refractive index of the semiconductor material ensures that the reflectance at the material/air interface is sufficiently large even though it is only about 0.32.

EXAMPLE 5.4 Reflectance at a GaAs/air interface

We may confirm that the reflectance at a GaAs surface is quite high by virtue of the high refractive index ($= 3.6$) of GaAs. From the Fresnel equations (section 8.1) we have

$$R = \left(\frac{n_2 - n_1}{n_2 + n_1} \right)^2 = \left(\frac{3.6 - 1}{3.6 + 1} \right)^2 \approx 0.32$$

The diode is cleaved along natural crystal planes normal to the plane of the junction so that the end faces are parallel; no further treatment of the cleaved faces is usually necessary, although occasionally optical coatings are added for various purposes. For GaAs, the junction plane is (100) and the cleaved faces are (110) planes.

The radiation generated within the active region spreads out into the surrounding lossy GaAs, although there is, in fact, some confinement of the radiation within a region called

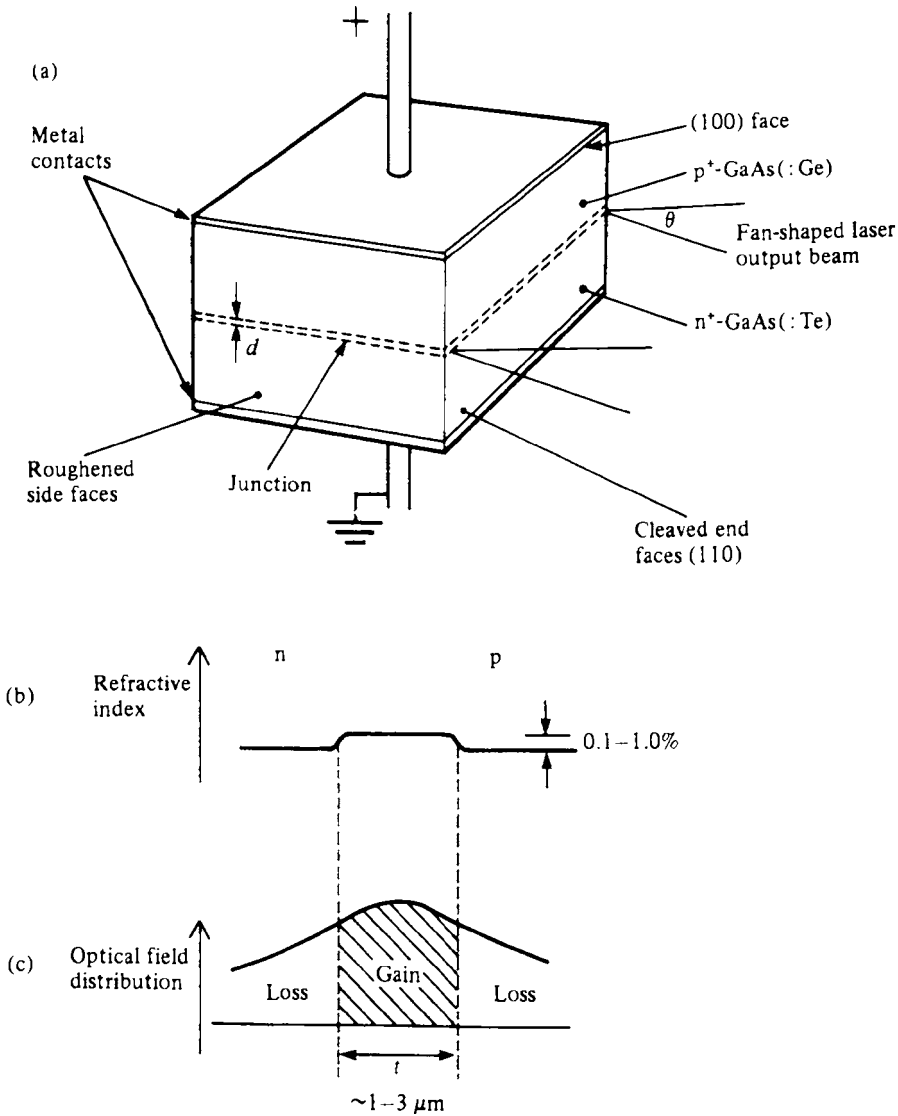


FIG. 5.24 Schematic construction of a GaAs homojunction semiconductor diode laser having side lengths 200–400 μm (a). The emission is confined to the junction region. The narrow thickness d of this region causes a large beam divergence. The very small change in refractive index in the junction region is shown in (b) and (c) shows the resulting poor confinement of the optical radiation to the gain region.

the *mode volume* (Fig. 5.23). The additional carriers present in the active region increase its refractive index above that of the surrounding material, thereby forming a dielectric waveguide (see Chapter 8). As the difference in refractive index between the centre waveguiding layer and the neighbouring regions is only about 0.02, the waveguiding is very inefficient and the radiation extends some way beyond the active region, thereby forming the mode volume. The waveguiding achieved in simple homojunction laser diodes of the form shown in Fig. 5.24 only works just well enough to allow laser action to occur as a result of very vigorous pumping. Indeed homojunction lasers can usually only be operated in the pulsed mode at room temperature because the threshold pumping current density (see below) required is so high, being typically of the order of 400 A mm^{-2} .

The onset of laser action at the threshold current density is detected by an abrupt increase in the radiance of the emitting region, as shown in Fig. 5.25, which is accompanied by a dramatic narrowing of the spectral width of emission. This is illustrated very clearly in Fig. 5.26 which shows the mode structure below, and at threshold, where the energy has been channelled into a relatively small number of modes. If the current is increased substantially above threshold one mode usually predominates, with a further decrease in the spectral width of the emission.

5.10.2.1 Threshold current density for semiconductor lasers

An exact calculation of the threshold current for a semiconductor laser is complicated by the difficulty of defining what is meant by a population inversion between two *bands* of energy levels. To simplify the problem, however, and to gain some insight into the important factors, we use the idealized structure shown in Fig. 5.23. We let the active volume, where population inversion is maintained, have thickness t and the mode volume, where the generated electromagnetic mode is confined, be of thickness d ($d > t$). In other lasers, the mode volume is usually smaller than the volume within which population inversion is maintained.

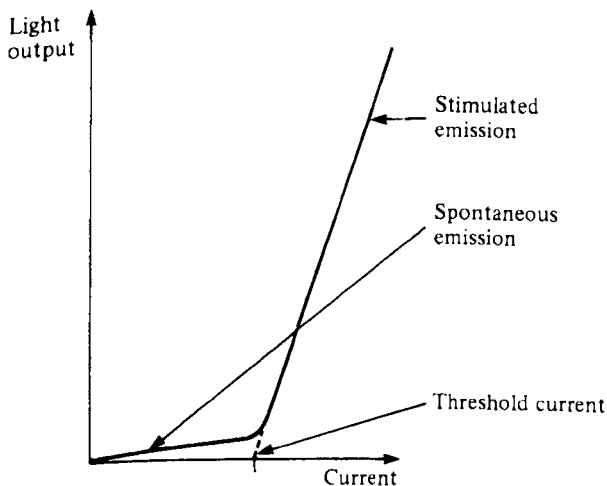


FIG. 5.25 Light output–current characteristic of an ideal semiconductor laser.

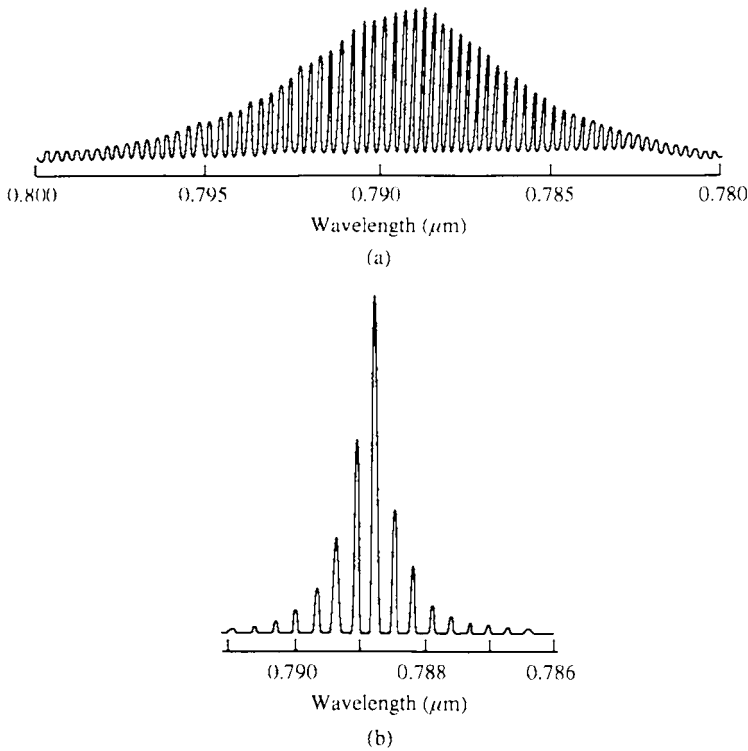


FIG. 5.26 Emission spectrum of a GaAlAs laser diode both just below (a) and just above (b) threshold. Below threshold a large number of Fabry-Perot cavity resonances can be seen extending across a wide LED-type spectrum. Above threshold only a few modes close to the peak of the gain curve oscillate. For the particular laser shown here the threshold current was 37 mA while spectra (a) and (b) were taken with currents of 35 mA and 39 mA, respectively.

A consequence of the situation in semiconductor lasers is that the portions of the mode propagating outside the active region may be absorbed. This offsets to some extent the gain resulting from those parts of the mode propagating within the active region. We allow for this by assuming that the effective population inversion within the mode volume ($d \times l \times w$) is given by reducing the actual population inversion in the active region by the factor t/d .

Referring to eq. (5.23), the threshold condition will thus be reached when

$$N_{\text{th}} = \left(N_2 - \frac{g_2}{g_1} N_1 \right)_{\text{th}} = \frac{d}{t} \left(\frac{8\pi\nu_0^2 k_{\text{th}} \tau_{21} \Delta\nu n^2}{c^2} \right)$$

We next assume that within the active region we can ignore N_1 , that is there is a large number of holes in the valence band; hence,

$$(N_2)_{\text{th}} = \frac{d}{t} \left(\frac{8\pi\nu_0^2 k_{\text{th}} \tau_{21} \Delta\nu n^2}{c^2} \right) \quad (5.38)$$

If the current density flowing through the laser diode is $J \text{ A m}^{-2}$, then the number of electrons per second being injected into a volume t (i.e. a region of thickness t and of unit cross-sectional area) of the active region is J/e . Thus the number density of electrons being injected per second is J/et electrons $\text{s}^{-1} \text{ m}^{-3}$. The equilibrium number density of electrons in the conduction band required to give a recombination rate equal to this injection rate is N_2/τ_c , where τ_c is the electron lifetime (τ_c is not necessarily equal to τ_{21} , the spontaneous lifetime, since some non-radiative recombination mechanisms are likely to be present).

The threshold current density is then given by

$$\frac{(J)_{\text{th}}}{et} = \frac{(N_2)_{\text{th}}}{\tau_c}$$

and substituting from eq. (5.38) we have

$$(J)_{\text{th}} = \frac{et}{\tau_c} \frac{d}{t} \frac{(8\pi\nu_0^2 k_{\text{th}} \tau_{21} \Delta\nu n^2)}{c^2}$$

EXAMPLE 5.5 Threshold current density in a GaAs laser

We may use the following data to estimate the threshold current density of a GaAs junction laser: wavelength, $\lambda = 0.84 \text{ }\mu\text{m}$; transition linewidth, $\Delta\nu = 1.45 \times 10^{13} \text{ Hz}$; loss coefficient, $\gamma = 3.5 \times 10^3 \text{ m}^{-1}$; refractive index, $n = 3.6$; dimensions, $l = 300 \text{ }\mu\text{m}$, $d = 2 \text{ }\mu\text{m}$; and internal quantum efficiency, $\eta_i \approx 1$.

Taking $n = 3.6$ gives $R = 0.32$. The threshold gain is given by eq. (5.17) as

$$(k)_{\text{th}} = \gamma + \frac{1}{2l} \ln \left(\frac{1}{R_1 R_2} \right)$$

Therefore

$$(k)_{\text{th}} = 3500 + \frac{1}{8 \times 10^{-4}} \ln \left(\frac{1}{0.32} \right)^2$$

that is

$$(k)_{\text{th}} = 7298 \text{ m}^{-1}$$

Hence, from eq. (5.39),

$$(J)_{\text{th}} = 15.5 \times 10^6 \text{ A m}^{-2} = 15.5 \text{ A mm}^{-2}$$

This value is in reasonable agreement with those measured at low temperature in GaAs lasers.

Substituting for k_{th} from eq. (5.17) then gives

$$(J)_{\text{th}} = \frac{8\pi\nu_0^2 ed \tau_{21} \Delta\nu n^2}{\tau_c c^2} \left[\gamma + \frac{1}{2l} \ln \left(\frac{1}{R_1 R_2} \right) \right] \quad (5.39)$$

The ratio τ_e/τ_{21} in this equation is often written as η_i , the internal quantum efficiency, which is the fraction of the injected electrons (or holes) which recombine radiatively.

5.10.2.2 Power output of semiconductor lasers

A discussion of power output and saturation in semiconducting lasers is basically the same as that for other lasers given in section 5.8. As the injection current increases above threshold, laser oscillations build up and the resulting stimulated emission reduces the population inversion until it is clamped at the threshold value. We can then express the power emitted by stimulated emission as

$$P = A[J - (J)_{th}] \frac{\eta_i h\nu}{e}$$

where A is the junction area.

Part of this power is dissipated inside the laser cavity and the rest is coupled out via the end crystal faces. These two components are proportional to γ and $(1/2l) \ln(1/R_1 R_2)$ respectively. Hence we can write the output power as

$$P_0 = \frac{A[J - (J)_{th}]\eta_i h\nu}{e} \frac{[(1/2l)\ln(1/R_1 R_2)]}{\gamma + (1/2l)\ln(1/R_1 R_2)} \quad (5.40)$$

The external differential quantum efficiency η_{ex} is defined as the ratio of the increase in photon output rate resulting from an increase in the injection rate (i.e. carriers per second), that is

$$\eta_{ex} = \frac{d(P_0/h\nu)}{d\{(A/e)[J - (J)_{th}]\}}$$

From eq. (5.40) we can write η_{ex} as

$$\eta_{ex} = \eta_i \left(\frac{\ln(1/R_1)}{\gamma l + \ln(1/R_1)} \right) \quad (5.41)$$

assuming that $R_1 = R_2$. Equation (5.41) enables us to determine the internal quantum efficiency from the experimentally measured dependence of η_{ex} on l ; η_i in GaAs is usually in the range 0.7–1.0.

Now if the forward bias voltage applied to the laser is V_f , then the power input is $V_f AJ$ and the efficiency of the laser in converting electrical input to laser output is

$$\eta = \frac{P_0}{V_f AJ} = \eta_i \left(\frac{J - (J)_{th}}{J} \right) \left(\frac{h\nu}{eV_f} \right) \frac{\ln(1/R_1)}{\gamma l + \ln(1/R_1)} \quad (5.42)$$

From Fig. 5.18, $eV_f \approx h\nu$ and therefore, well above threshold ($J \gg (J)_{th}$) where optimum coupling ensures that $(1/l) \ln(1/R_1) \gg \gamma$, η approaches η_i . As noted above, η_i is high (~ 0.7) and thus semiconductor lasers have a very high power efficiency.

5.10.2.3 Heterojunction lasers

As we noted above, the threshold current density for homojunction lasers is very large owing to poor optical and carrier confinement, which results in the parameters d and γ in eq. (5.39) being large. Dramatic reductions in the threshold current density to values of the order of 10 A mm^{-2} at room temperature coupled with higher efficiency can be achieved using lasers containing heterojunctions (see section 2.8.6 and ref. 5.16). The properties of heterostructure lasers which permit a low threshold current density and CW operation at room temperature can be illustrated with the double heterostructure (DH) laser illustrated in Fig. 5.27. In this structure, a layer of GaAs, for example, is sandwiched between two layers of the ternary compound $\text{Ga}_{1-x}\text{Al}_x\text{As}$ which has a wider energy gap than GaAs and also a lower refractive index. Both N-n-P and N-p-P structures show the same behaviour (where N and P represent the wider bandgap semiconductor, according to carrier type).

Figure 5.27(b) also shows that carrier and optical confinement may be achieved simultaneously. The bandgap differences form potential barriers in both the conduction and valence bands which prevent electrons and holes injected into the GaAs layer from diffusing away. The GaAs layer thus becomes the active region, and it can be made very narrow so that t is very small, typically about $0.2 \mu\text{m}$. Similarly, the step change in refractive index provides a very much more efficient waveguide structure than was the case in homojunction lasers. The radiation is therefore confined mainly to the active region. In addition, the fraction of the propagating mode which lies outside the active region is in a wider bandgap semiconductor and is therefore not absorbed, so that γ is much smaller than in homojunction lasers.

Further reductions in threshold current can be obtained by restricting the current along the junction plane into a narrow 'stripe' which may only be a few micrometres wide. Such stripe geometry lasers have been prepared in a variety of different ways; typical examples are shown in Fig. 5.28. In Fig. 5.28(a), the stripe has been defined by proton bombardment of the adjacent regions to form highly resistive material, whereas in Fig. 5.28(b) a mesa structure has been formed by etching; an oxide mask prevents shorting of the junction during metallization to form contacts. With stripe geometry structures, operating currents of less than 50 mA can produce output powers of about 10 mW.

Stripe geometry devices have further advantages including the facts that (a) the radiation is emitted from a small area which simplifies the coupling of the radiation into optical fibers (see Chapter 9) and (b) the output is more stable than in other lasers. A close examination of typical light output-current characteristics reveals the presence of 'kinks' as shown in Fig. 5.29(a). These 'kinks' are associated with a sideways displacement of the radiating filament within the active region (the radiation is usually produced from narrow filaments within the active region rather than uniformly from the whole active region). This lateral instability is caused by interaction between the optical and carrier distributions which arises because the refractive index profile, and hence the waveguiding characteristic, is determined, to a certain extent, by the carrier distribution within the active region. The use of very narrow stripe regions limits the possible movement of the radiating filament and eliminates the 'kinks' in the light output-current characteristics as shown in Fig. 5.29(b). The structures shown in Fig. 5.28 are referred to as *gain guiding* because the width of the gain region is determined by the restriction of the extent of the current flow, which of course creates the

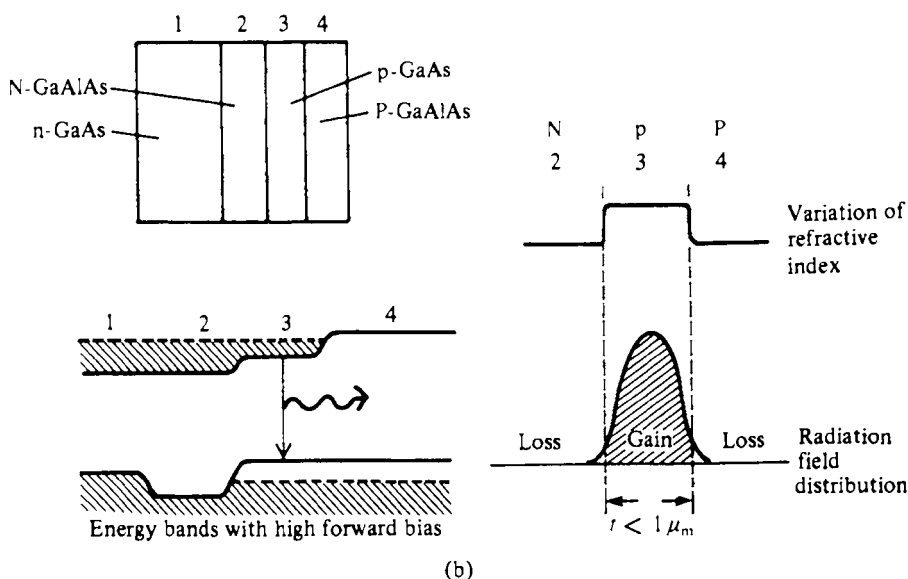
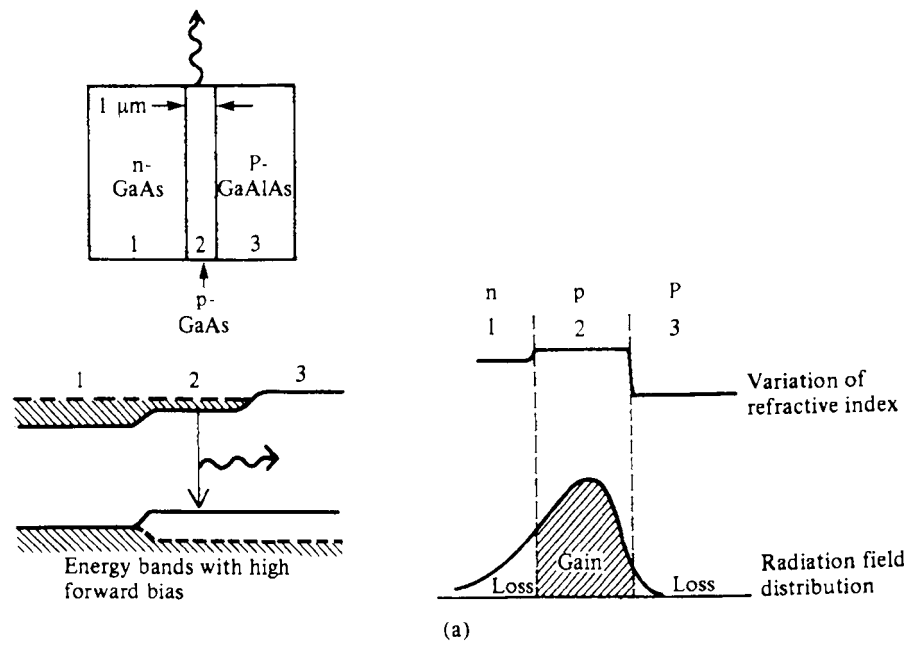


FIG. 5.27 Diagram illustrating the action of single (a) and double (b) heterojunction structures in confining the carriers and radiation to the gain region (as before, in the diagrams of the energy bands, the dashed lines represent the Fermi levels after forward bias has been applied).

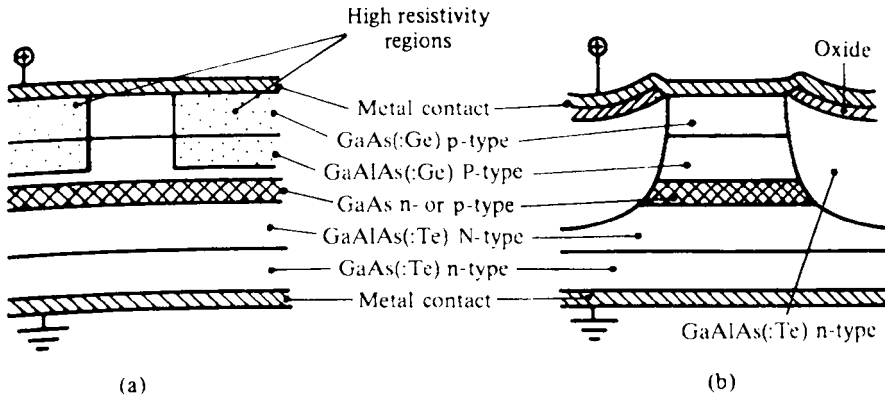


FIG. 5.28 Schematic cross-section (end view) of two typical stripe geometry laser diodes: (a) the stripe is defined by proton bombardment of selected regions to form high resistivity material; (b) the stripe is formed by etching a mesa structure and then GaAlAs is grown into the previously etched outsides of the active region to form a 'buried stripe' structure.

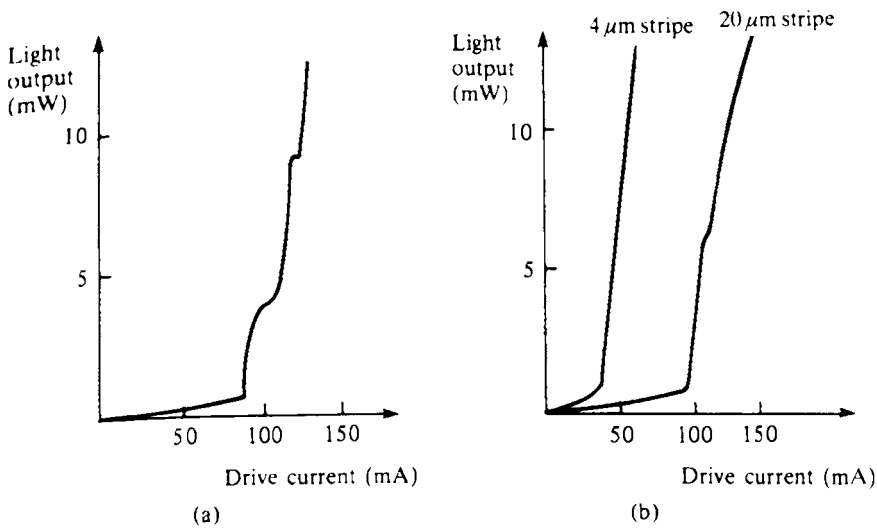


FIG. 5.29 Light output-current characteristics of (a) a laser showing a lateral instability or 'kink' and (b) stripe lasers, in which the 'kinks' have disappeared.

population inversion, and hence the gain, within the active region. Alternatively stripe geometry lasers can be fabricated using *index-guided* structures, in which an optical waveguide is created (section 8.2) as illustrated in Fig. 5.30(a). Here a *buried heterostructure* has been formed, in which the active region is surrounded on all four sides by material of lower refractive index. The creation of such structures in practice is quite complex; a relatively simple one is shown in Fig. 5.31(a). One relatively straightforward alternative is to change the thickness of the semiconductor layer next to the waveguide (Fig. 5.30b) which creates an effective refractive index difference between the active region and those next to it in the same

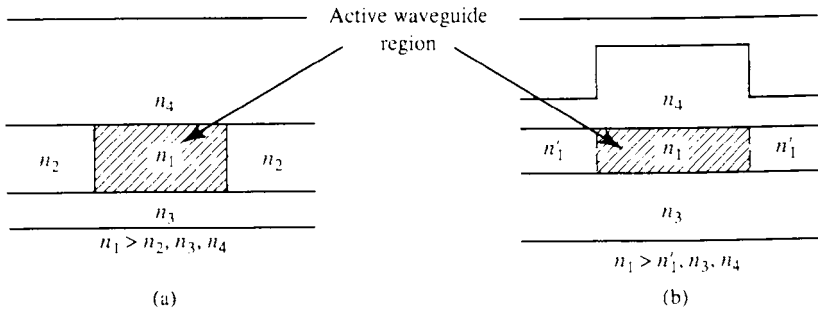


FIG. 5.30 Schematic representation of (a) a buried heterostructure which acts as a waveguide (end view), and (b) a structure which behaves like a buried heterostructure; the varying thickness of the layer next to the guiding layer creates changes in the apparent refractive index, thereby achieving a waveguiding structure.

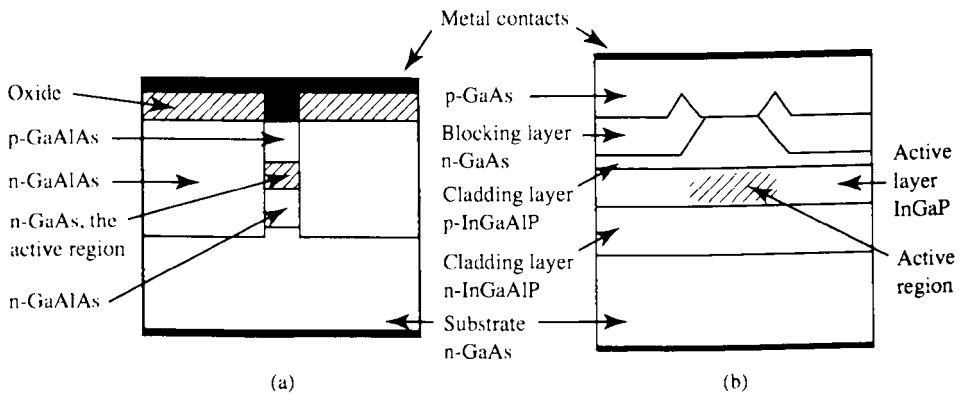


FIG. 5.31 Buried heterostructure index guiding laser structures: (a) based on InGaAsP (and the structure shown in Fig. 5.30a); (b) based on GaAs (and the structure shown in Fig. 5.30b).

layer. A device based on this technique is shown in Fig. 5.31(b). Several other buried layer heterostructure devices are described in ref. 5.17.

In general gain-guided lasers are easier to fabricate than index-guided lasers, but their poorer optical confinement limits the beam quality, and makes stable, single mode operation difficult to achieve. On the other hand the fact that the beam spread is greater reduces the optical power density at the output face thereby reducing the risk of damage (see below).

Although the application of diode lasers to fiber optical communications will be discussed in detail in Chapter 9 it is perhaps appropriate to mention at this stage one or two points of relevance. These include the temperature dependence of the threshold current, output beam spread, degradation and the use of materials other than GaAlAs.

The threshold current density J_{th} increases with temperature in all types of semiconductor laser but, as many factors contribute to the temperature variation, no single expression

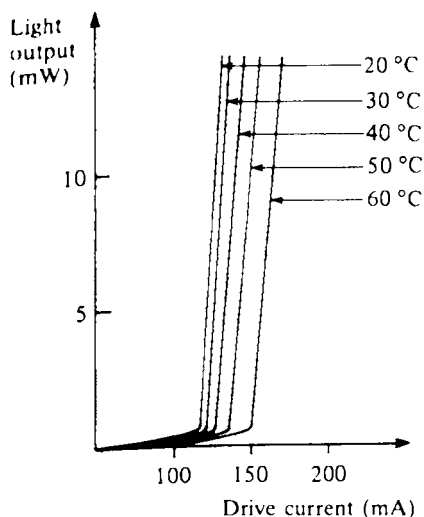


FIG. 5.32 Light output–current characteristics of a 20 μm stripe laser as a function of temperature.

is valid for all devices and temperature ranges. Above room temperature, which is usually the region of practical interest, it is found that the ratio of J_{th} at 70°C to J_{th} at 22°C for GaAlAs lasers is about 1.3–1.5 with the lowest temperature dependence occurring for an aluminium concentration such that the bandgap energy difference is 0.4 eV. Typical light output–current characteristics for a GaAlAs DH laser are shown in Fig. 5.32.

The angular spread of the output beam depends on the dimensions of the active region and the number of oscillating modes (which, in turn, depends on the dimensions of the active region, the refractive index and the pump power). For wide active regions, we find that the beam divergence both parallel to (θ_{\parallel}) and perpendicular to (θ_{\perp}) the plane of the junction is given approximately by simple diffraction theory. Thus, normal to the junction plane we have $\theta_{\perp} \approx 1.22\lambda/d$. For DH lasers, where the active region is much narrower, θ_{\perp} is given approximately by $\theta_{\perp} \approx 1.1 \times 10^{-3}x(t/\lambda)$, where x is the mole fraction of aluminium. Thus for a DH laser with $t = 0.1 \mu\text{m}$, $x = 0.3$ and $\lambda = 0.9 \mu\text{m}$, we find $\theta_{\perp} = 37^\circ$ (in good agreement with experimental observations).

Until recently, the system $\text{Ga}_{1-x}\text{Al}_x\text{As}/\text{GaAs}$ was the most widely investigated and used for the production of DH lasers. There are many reasons for this, including the facts that (a) GaAs is a direct bandgap semiconductor which can easily be doped n- or p-type; (b) the ternary compound $\text{Ga}_{1-x}\text{Al}_x\text{As}$ can be grown over a wide range of compositions, and not only does it have a very close lattice match to GaAs ($\approx 0.1\%$) for all values of x (thus there is low interfacial strain between adjacent layers and consequently very few strain-induced defects at which non-radiative recombination may occur), but it is also a direct bandgap semiconductor for $x < 0.45$; and (c) the relative refractive indices and bandgaps of GaAs and $\text{Ga}_{1-x}\text{Al}_x\text{As}$ provide for optical and carrier confinement.

In optical fiber communications, however, it is desirable to have a laser emitting at wavelengths in the region 1.1 to 1.6 μm where present optical fibers have minimum attenuation

and dispersion. Wavelengths in this range can be obtained from lasers fabricated from quaternary compounds such as $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$ because of the wide range of bandgaps and lattice constants spanned by this alloy. Figure 5.33 shows the lattice constant variation with bandgap (and emission wavelength) for this alloy. By suitable choice of x and y , exact lattice matching to an InP substrate can be achieved and strain-free heterojunction devices can be produced. The GaInAsP layers may be grown on InP substrates by liquid phase, vapour phase or molecular beam epitaxial methods (see ref. 5.18). A typical DH stripe contact laser diode of GaInAsP/InP emitting at $1.1\text{--}1.3\text{ }\mu\text{m}$ is shown schematically in Fig. 5.34.

The question of laser reliability is also important in relation to applications such as telecommunications. Laser life may be limited by ‘catastrophic’ or ‘gradual’ degradation. Catastrophic failure results from mechanical damage to the laser facets due to too great an optical flux density. The damage threshold is reduced by the presence of flaws on the facets; however, it may be increased by the application of half-wave coatings of materials such as Al_2O_3 . While facet damage is more likely in lasers operating in the pulse mode, it can also

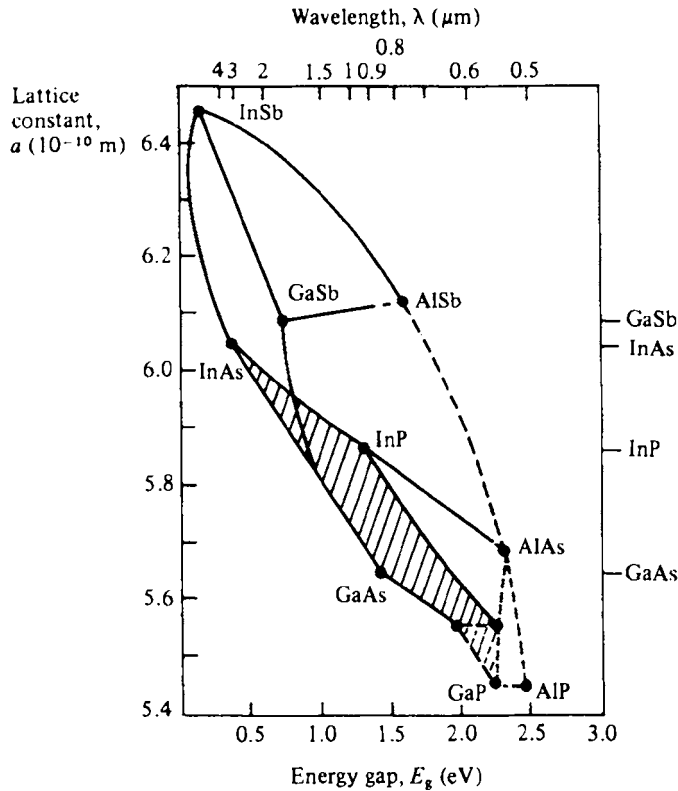


FIG. 5.33 Lattice constant versus energy gap for various III-V compounds. The solid lines correspond to direct bandgap materials, the dashed lines to indirect bandgap materials. The binary compounds which may be used for lattice-matched growth are indicated on the right. The shaded area shows the quaternary compound $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$. By adjusting x and y , lattice matching to both GaAs and InP is possible; the latter match, for example, occurs for $x = 0.8$, $y = 0.35$.

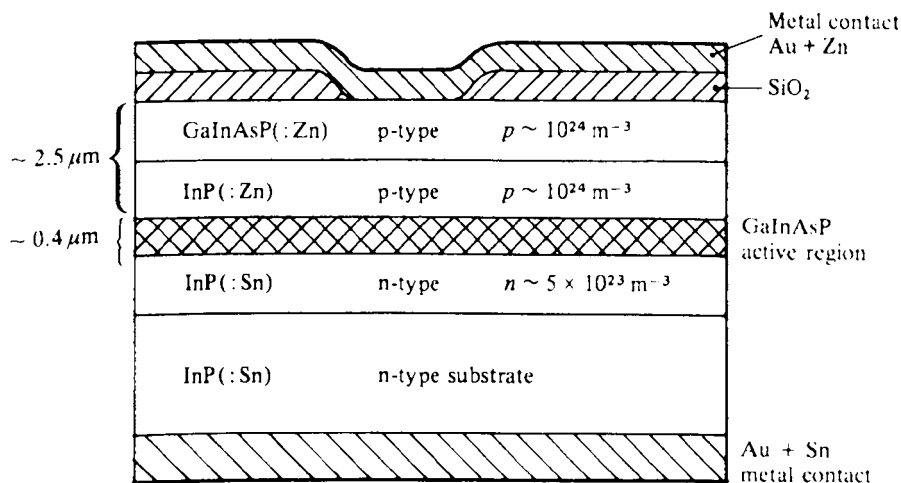


FIG. 5.34 Schematic diagram of a double heterojunction stripe contact laser diode of the quaternary compound $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$ on an InP substrate with (100) orientation.

occur in CW-operated lasers. This is so especially in the central portion of the active region of stripe lasers where the optical flux density is greatest. Uncoated lasers with stripes about $20 \mu\text{m}$ wide tend to fail catastrophically when the optical flux exceeds about 10^9 W m^{-2} .

Gradual degradation depends principally on the current density, but also on the duty cycle and fabrication process. It has been observed that as time elapses the threshold current density increases, 'dark' lines develop in the emission and then the CW output falls off drastically.

The development of dark lines is apparently related to the formation, in the vicinity of the active region, of so-called *dark-line defects*, which act as non-radiative recombination centres. Dark-line defects are attributed to defects such as dislocations, which may have a number of sources. These include (a) edge dislocations formed to relieve stress caused by interfacial lattice mismatch, (b) bonding of the laser to the heat sink and (c) impurities introduced during substrate preparation and heteroepitaxial wafer growth. Furthermore, the energy released by non-radiative electron-hole recombination may result in the creation and migration of defects.

Defects may be formed in the active region during device fabrication or penetrate into it during subsequent operation. Dark-line defects may grow owing to a process called dislocation climb (i.e. the movement of dislocations involving atomic transport to or away from the dislocation) and extend throughout the device structure.

Dislocation growth may be stimulated by carrier injection and recombination. GaAs lasers initially containing dislocations are found to degrade at a much higher rate than those that are initially dislocation free. Furthermore, devices with exposed edges that contain edge defects also degrade more rapidly than those in which recombination is restricted to internal regions of the crystal.

Thus, to produce lasers with long lifetimes great care must be taken with substrate selection, and wafer processing and crystal growth must be carried out under ultraclean condi-

tions to fabricate a laser with a strain-free structure. Despite these problems, lasers with lifetimes in excess of 40 000 hours are now available corresponding to continuous operation over a 5 year period.

5.10.2.4 Quantum well lasers

Quantum well structures were discussed in section 2.9, where it was shown that in very narrow semiconductor layers (i.e. the quantum wells) there is a very significant increase in the density of states near the bottom of the conduction band and the top of the valence band. The increased densities of states enable a population inversion to be obtained more easily and, as a consequence of this, and the very small active volume, the threshold currents in quantum well lasers are about a factor of 10 less than those in DH lasers. In addition, quantum well lasers have low temperature sensitivity and their output characteristics are free from kinks. Such lasers are therefore increasingly replacing DH lasers as materials growth technology improves enabling the controlled fabrication of very thin structures in an increasingly wide range of semiconductors (ref. 5.19).

One of the problems with the single quantum well (SQW) structure described above is that, because of the extreme narrowness of the active region, optical confinement is very poor. This causes higher losses and tends to negate the potential advantages of low threshold currents. One way of reducing these problems is to use the multiple quantum well (MQW) structure illustrated in Fig. 5.35(b), which because of its greater thickness gives better optical confinement and beam definition.

The single quantum well can be extended to coupled quantum wells, to form the MQW laser, Figs 5.35(a) and (b). In such devices several GaAs quantum wells, for example, may be coupled by very thin intervening GaAlAs *barrier* layers. The overall active region is now thicker so that the carriers which are not captured and therefore able to recombine in the first well may be captured by the second or a subsequent well. Although MQW lasers have larger threshold currents than single quantum well lasers, where the threshold current may be as low as 1 mA or less, they can emit more optical power, and their structure results in better optical confinement.

Further improvement in both optical and carrier confinement can be obtained by adding cladding layers and *separate confinement heterostructure* (SCH) layers as illustrated in Fig. 5.35(b) (ref. 5.19b). The SCH layers are chosen to have a refractive index which is greater than that of the cladding layers, so that total internal reflection occurs at the boundary. The SCH layers also, together with the barrier layers, have an energy gap, E_g , between that of the cladding layers and quantum wells, so that the charge carriers are confined between the cladding regions – hence the SCH layers are so named because the carriers and photons are separately confined.

The cladding layers are doped n- and p-type, while the MQW layers are undoped. Under forward bias the electrons and holes are injected from the cladding layers, diffuse across the SCH layers and enter the MQW structure where they recombine. The cavity mirrors are provided by the high reflectance of the device faces. Alternatively multilayer coatings can be added or the techniques described in section 6.2 can be used.

The lasing region of the active layer can be restricted to a narrow strip thereby in effect

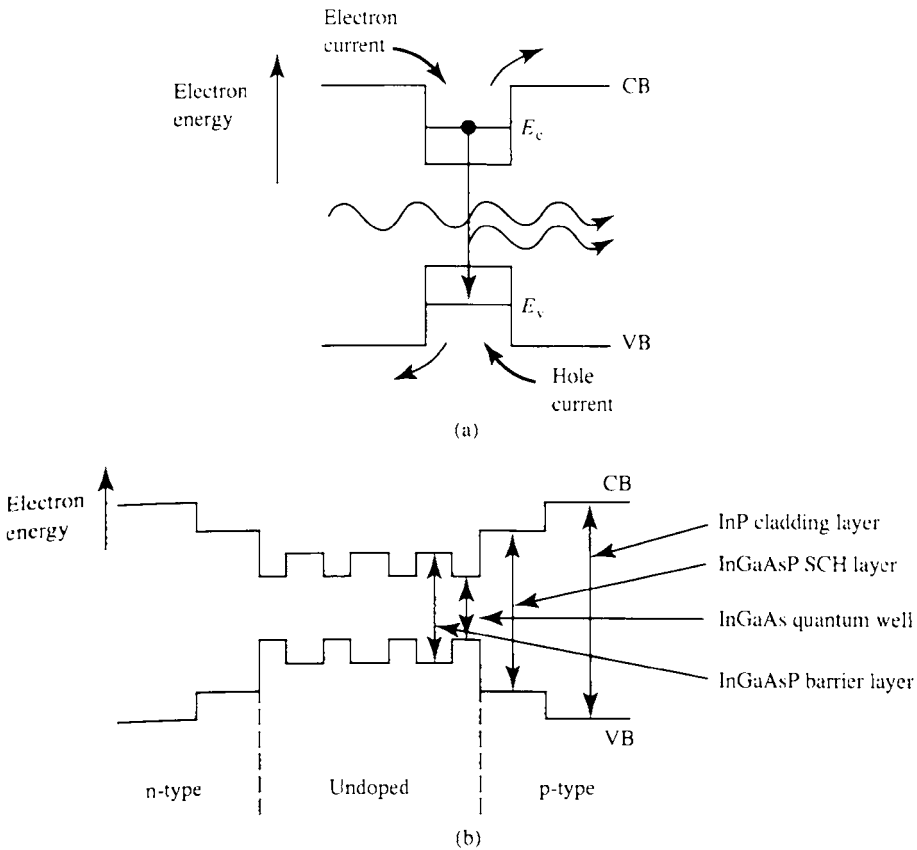


FIG. 5.35 (a) Stimulated emission in a single quantum well. Some of the carriers escape (thin arrows) without contributing to the process. (b) The energy band diagram for a typical multiple quantum well (MQW) laser with separate confinement heterostructure (SCH) layers. The light waves are transversely guided between the cladding layers, either by changes in the refractive index (index guiding), or by current injection from a stripe contact (gain guiding). By using InGaAsP, InGaAs or InGaAlAs semiconductor systems grown on InP substrates the wavelength range from 920 nm to 1650 nm can be covered.

confining the carriers in two dimensions. Such structures are referred to as *quantum wire microcavities*, and are the basis of QWR-MC lasers. Further restriction, that is into three dimensions, gives rise to *quantum dot* lasers. Despite manufacturing difficulties quantum wire and quantum dot arrays are potentially important because, in addition to very low threshold currents, they have very high modulation bandwidths, narrow spectral linewidths and low temperature sensitivity.

5.10.2.5 Arrays – vertical cavity lasers

The output power from semiconductor lasers may be increased by using one-dimensional arrays of single mode lasers on a bar of semiconductor as shown in Fig 5.36(a). Such arrays

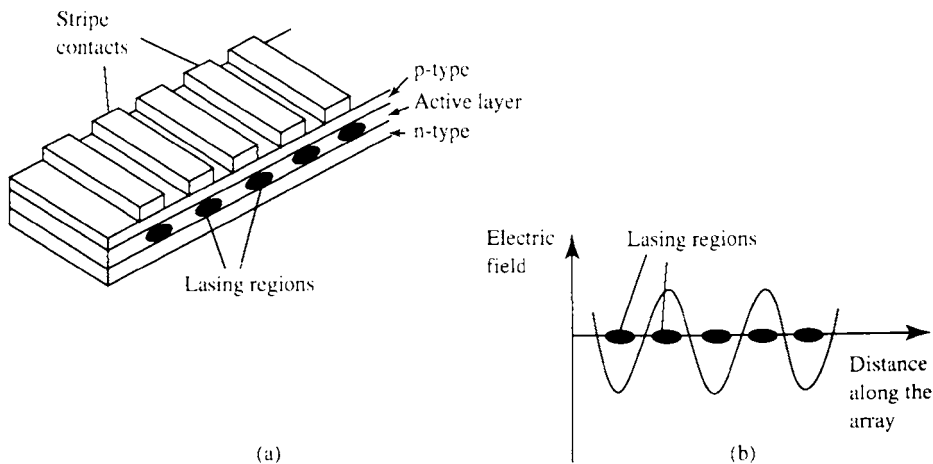


FIG. 5.36 A linear array of lasing elements formed within a single semiconductor bar: (a) shows the stripe contacts, which define the lasing regions; (b) shows the electric field distribution; the field is zero midway between the lasing elements where there is absorption rather than gain.

are called *phased* arrays since the electric fields associated with the individual elements interact with each other resulting in definite phase relationships between them. Frequently the phase difference between adjacent elements is arranged to be 180° (Fig. 5.36b), so that the resultant field midway between the active regions of adjacent elements is zero. These midway regions are more likely to exhibit absorption rather than gain so the overall losses are minimized. Unfortunately, this phase relationship results in a power distribution in the plane of the active layer with an angular distribution comprising two lobes rather than a single one. In fact a single-lobed power output distribution can be achieved if the phase between adjacent elements is zero. The phase difference can be controlled by a number of techniques including variation of the lateral spacing between the elements in the array (ref. 5.20).

Linear arrays are available in widths up to 10 mm and can generate CW powers up to 20 W. Outputs of 10 kW or more can be achieved by stacking up to 200 linear array bars together. It is important to realize that as the power output increases, so too do the cooling requirements; it is therefore vital to consider carefully how to remove excess heat to prevent the array from self-destructing. Very high power arrays, for example, require water cooling.

VERTICAL CAVITY LASERS

A structure which particularly lends itself to the fabrication of laser arrays is the *vertical cavity* surface-emitting laser (VCSEL) (ref. 5.20). While in traditional, horizontal edge-emitting lasers the resonant cavity is in the plane of the active layer, in VCSELs (Fig. 5.37) it is perpendicular to this plane. The light resonates between mirrors on the top and bottom of the laser wafer so that the photons pass through only a very short length (typically $\leq 1 \mu\text{m}$) of active medium, in which they can stimulate emission. Thus VCSELs have very much lower round trip gain than horizontal edge-emitting lasers, and consequently require highly reflecting mirrors (reflectance ≥ 0.9) to sustain oscillations. Clearly the reflectance of the semi-

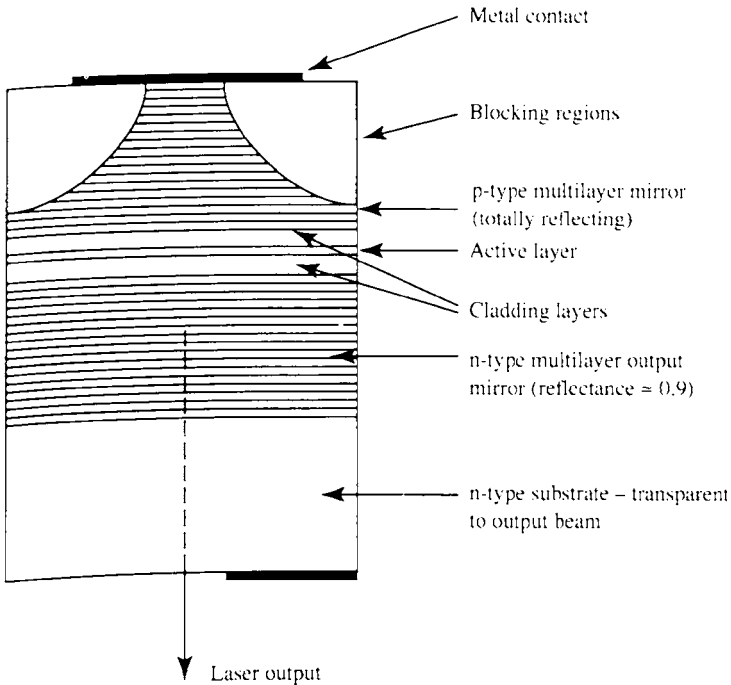


FIG. 5.37 A vertical cavity, surface-emitting laser (VCSEL).

conductor facets at about 0.32 is insufficient and multilayer mirrors comprising several tens of alternate $\lambda/4$ coatings of AlAs and AlGaAs are often used as illustrated in Fig. 5.37 (ref. 5.20f).

The active layer comprises an SQW or MQW structure, which together with cladding and confinement layers forms an optical cavity which is one wavelength thick. The active region is arranged to be at the peak of the standing wave formed between the mirrors.

All vertical cavity lasers emit from their surface rather than their edge (though surface-emitting lasers are available which do not have vertical cavities). The emission is typically from round or square areas which are about $10\ \mu\text{m}$ wide so that the output beams are highly symmetrical in contrast to those of edge-emitting devices. Divergence angles are only 7° to 10° , and by using microlenses integrated onto the device surface some 90% of the output may be coupled into optical fibers.

In addition to the symmetrical beam profile, low threshold currents and good temperature stability of VCSELs, a major attraction of surface emission is the ability to fabricate monolithic one- and two-dimensional arrays of many elements. In practice it is possible to grow many thousands of VCSELs simultaneously on a 3 inch (75 mm) wafer and, equally importantly in relation to manufacturing costs, to test these and measure the optical and electrical properties *in situ*.

A range of one-dimensional (up to 64 elements) and two-dimensional (8×8) VCSEL arrays is now commercially available, with much larger arrays under development. Each laser in the array can be independently addressed so that, for example, the lasers in an array can

act as sources for several parallel communication channels (ref. 5.20b), particularly as vertical cavity lasers have very high modulation bandwidths.

VCSELs are currently available in the wavelength range 650–690 nm using GaAs/GaAlAs and 850–980 nm using InGaAs/GaAs semiconductor systems. Unfortunately efforts to fabricate VCSELs operating CW at room temperature in the wavelength range 1300–1550 nm, which is so important for long range fiber optic communications, have not yet succeeded.

5.10.2.6 Short wavelength lasers

Recently there has been increased demand for shorter wavelength semiconductor lasers for applications such as compact disc and optical storage, colour printing and semiconductor lithography. The shorter the wavelength the smaller is the area of a focused beam ($\sim \lambda^2$) thereby allowing increased storage capacity, and similarly the narrower the features than can be created with optical lithography.

Recently red lasers based on AlGaInP have become available for use in barcode readers, while quantum well lasers with GaInP active layers have enabled wavelengths as short as 630 nm to be generated.

Despite the improved reliability of semiconductor lasers emitting in the red and yellow parts of the spectrum, reliable lasers emitting in the green and blue remain elusive. However, recent improvements in materials technology have enabled CW, room temperature operation to be demonstrated in so-called II–VI semiconductors such as ZnSe, ZnMgSSe and related compounds on GaAs substrates. Alternatively CW laser operation at a wavelength of 417 nm has been obtained from devices based on gallium nitride (GaN) which is a rather difficult material to work with (ref. 5.21a). These lasers contain an MQW structure of 26 quantum wells 2.5 nm thick of $\text{In}_{0.2}\text{Ga}_{0.8}\text{N}$ separated by layers of $\text{In}_{0.05}\text{Ga}_{0.95}\text{N}$ barriers 5.0 nm thick giving a total thickness of some 200 nm. The threshold current densities and operating voltages are still rather high at about 10 kA cm^{-2} and 25 V respectively, but these are being steadily reduced as the technology develops.

The requirement of close lattice matching (i.e. $\approx 0.1\%$) for the components in a heterojunction structure made it difficult to cover some wavelength ranges. A recent development which has helped in this respect is the discovery that very thin layers (less than a few tens of nanometres) can accommodate a lattice mismatch of more than 1%. These layers are called *strained* lattice layers, and the technique was first used to enable the fabrication of InGaAs/GaAs lasers emitting at 980 nm. Strained layers are also used in quantum well structures to produce active layers which need not be precisely matched to the surrounding layers. This technique was used to produce the lasers based on ZnSe, which emit in the green at a wavelength of 525 nm (ref. 5.21b), and to enable GaN to be grown on mismatched substrates such as sapphire, which has the same crystal structure, or silicon nitride.

5.10.2.7 Superluminescent light-emitting diodes

We end this section with a discussion of a device which, while not a laser, does depend on optical amplification, namely the superluminescent light-emitting diode (SLD). SLDs have a structure which is rather similar to that of conventional injection laser diodes and edge-

emitting LEDs (ref. 5.22); indeed the SLD has optical properties which are intermediate between these two devices. Both stripe geometry and buried heterostructure SLDs are available, emitting at a range of wavelengths. In contrast to laser diodes, however, the non-output end of the device is made optically lossy to minimize feedback and suppress laser oscillations. This can be achieved simply by roughening the cleaved surface of the device to scatter the light, or by adding an antireflection coating.

In operation the injection current is increased until stimulated emission, and hence amplification of spontaneous emission, just occurs. That is, operation is on the 'knee' of the laser diode output characteristic shown in Fig. 5.25. Although there are no oscillations the stimulated emission, within a single pass through the device, provides gain so that the device output increases rapidly with increase in current – this is termed *superradiance* or *superluminescence*. High optical output power can be obtained together with a narrowing of the spectral width, which also results from the stimulated emission.

These characteristics of the output from SLDs give a number of advantages over conventional LEDs in relation to their use in fiber optic communications. These include: higher power outputs (up to 60–100 mW), a more directional light beam, and a narrower spectral linewidth, all of which improve the source to fiber coupling. Moreover, the superradiant emission process within SLDs tends to increase their modulation bandwidth. In contrast to conventional LEDs, however, SLDs suffer from having a non-linear output characteristic and an increased temperature dependence of the output power. Compared with laser diodes they require substantially higher injection currents (by a factor of about three) to produce a similar power output.

5.10.3 Gas lasers

Gas lasers are the most widely used type of laser; they range from the low power helium–neon (He–Ne) laser commonly found in teaching laboratories to very high power carbon dioxide lasers, which have many industrial applications. Basically, there are three different classes of gas laser, according to whether the transitions are between the electronic energy levels of atoms, or of ions, or between the vibrational/rotational levels of molecules. In general, the energy levels involved in the lasing process are well defined and the absence of broad bands effectively eliminates the possibility of optical pumping. Though other methods can be used, most gas lasers are excited by electron collisions in a gas discharge. We shall now consider a typical example from each of those classes mentioned above.

5.10.3.1 Atomic lasers – the He–Ne laser

In the He–Ne laser the active medium is a mixture of about 10 parts of helium to one part of neon. The neon provides the energy levels for the laser transitions (about 150 different laser transitions have been observed, although only the four shown in Fig. 5.38 are reasonably strong), while the helium atoms, though not directly involved in the laser transitions, have an important role in providing an efficient excitation mechanism for the neon atoms. Excitation usually takes place in a d.c. discharge created by applying a high voltage (≈ 2 to 4 kV) across the gas contained in a narrow diameter glass tube at a pressure of about 10 torr,

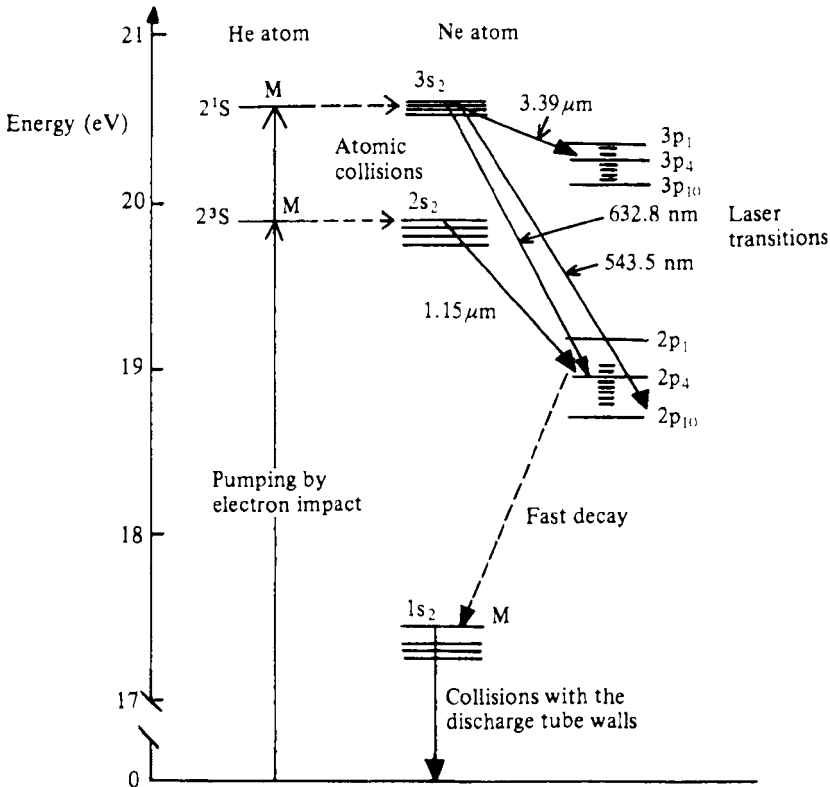
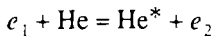


FIG. 5.38 Energy levels relevant to the operation of the He-Ne laser. M indicates a metastable state.

as illustrated in Fig. 5.39. As the discharge tube exhibits a negative dynamic resistance when a discharge has been initiated, it is necessary to include a load resistor to limit the current and protect the power supply.

The pumping process can be described as follows. The first step is the electron impact excitation of helium atoms to one of two metastable states designated 2¹S and 2³S; this is represented by



where e_1 and e_2 are the electron energies before and after the collision. While in one of their excited states (He^*), helium atoms can transfer their energy to neon atoms, with which they may collide. The probability of this *resonant transfer* of energy is proportional to $\exp(-\Delta E/kT)$ where ΔE is the energy difference between the excited states of the two atoms involved. The energy level diagram for helium and neon is shown in Fig. 5.38. Here, the neon states are labelled by the so-called Paschen notation (the numerical subscripts have no spectroscopic significance and are used only as labels) while the helium states are labelled according to the $L-S$ representation mentioned in section 5.10.1. Figure 5.38 shows that there is a group of four neon levels at almost the same energies as each of the two excited helium

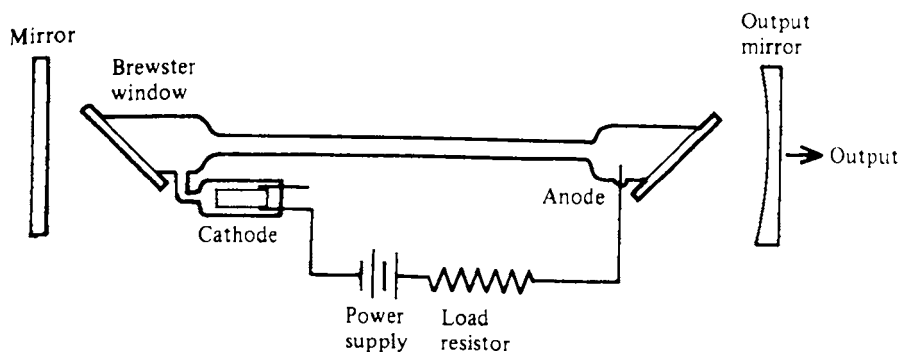
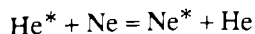


FIG. 5.39 Construction of a typical low power laser such as the He-Ne laser. The load resistor serves to limit the current once the discharge has been initiated.

states, and resonant transfer thus occurs quite readily. The energy transfer is represented by



A population inversion is thus created between the 3s and (3p, 2p) groups of levels and also between the 2s and 2p levels. Transitions between the 3s and 2s levels and between the 3p and 2p levels are forbidden by quantum mechanical selection rules.

The He-Ne laser is another example of a four-level system, and as such we require that the population of the lasing transition terminal level be kept as low as possible. This implies that electrons in the terminal level should decay as rapidly as possible back to the ground state. In neon this is a two-step process: the first, 2p to 1s, is a rapid transition, but the second, 1s to the ground state, is not so rapid. The latter transition rate is, however, enhanced by collisions with the walls of the discharge tube, and indeed the gain of the laser is found to be inversely proportional to the tube radius. For this reason, the discharge tube diameter should be kept as narrow as possible.

The transition 2p to 1s is also of interest since it gives rise to the familiar colour of 'neon lights'. Thus the 2p level itself must be populated by the discharge. This is unfortunate since an increase in the population of the 2p level implies a decrease in population inversion (at least as far as the 1.15 μm , 632.8 nm and 543.5 nm wavelengths are concerned), and in fact this effect is to a large extent responsible for lasing ceasing at high tube currents. We cannot therefore increase the output simply by increasing the current indefinitely. Thus the He-Ne laser is destined to remain a relatively low power device.

EXAMPLE 5.6 Efficiency of an He-Ne laser

We may estimate the efficiency of a low power He-Ne laser from the following.

A typical laser operates with a current of 10 mA at a d.c. voltage of 2500 V and gives an optical output of 5 mW. Its overall power efficiency is then

$$\frac{5 \times 10^{-3}}{2500 \times 1 \times 10^{-2}} = 0.02\%$$

Three of the four main laser transitions (at 3.39 μm , 632.8 nm and 543.5 nm) share a common starting level. Thus the transitions are always competing with each other and precautions must be taken to prevent the two unwanted wavelengths from lasing. This can be achieved quite easily by using the multilayer coated mirrors discussed in section 5.5, which have a wavelength-dependent reflectance. The very low absorption loss of such mirrors is an essential feature as the gain in the He–Ne medium is rather small; indeed, the use of such mirrors is quite general in gas lasers.

The basic structure of the He–Ne laser is relatively simple. The essential elements are shown in Fig. 5.39. The discharge is usually initiated by a high voltage ‘trigger’ pulse of some 10–20 kV, and then maintained at a current of 5 to 10 mA. The mirrors forming the resonant cavity are sometimes cemented to the ends of the discharge tube, thereby forming a gas-tight seal. Alternatively, the mirrors can be external to the tube, which is then sealed with glass windows orientated at the Brewster angle to the axis of the tube. This arrangement allows 100% transmission for the radiation with its electric vector vibrating parallel to the plane of incidence, thereby ensuring the maximum possible gain (minimum losses) in each round trip. The Brewster windows therefore also result in the output being plane polarized. Although this arrangement is slightly more complicated than the former one, it enables us to insert frequency stabilizing, mode selecting and other devices into the cavity. The mirrors can also be changed to allow operation with other output characteristics and at other wavelengths.

The power output from He–Ne lasers is rather small (up to about 100 mW maximum, though more typically a few milliwatts); however, the radiation is extremely useful in a wide range of applications because it is highly collimated, coherent and has an extremely narrow linewidth.

5.10.3.2 Ion lasers

NOBLE GAS ION LASERS

The most powerful CW lasers operating in the visible region are the inert gas ion lasers such as the argon and krypton ion lasers. CW outputs of several watts can readily be obtained while, if the laser is pulsed, powers up to a kilowatt in microsecond pulses can be generated.

The gas atoms are ionized by electron collision in a high current discharge (≈ 15 to 50 A). The ions are excited by further electron collisions up to a group of energy levels (4p) some 35 eV above the atomic ground state. As the electron energies are only a few electron volts, the excitation must be the result of multiple collisions. A population inversion forms between the 4p levels and the 4s level, which is about 33.5 eV above the ground state (Fig. 5.40), so that a series of stimulated lines is emitted. These range in wavelength from 351 nm to 520 nm, although most of the energy is concentrated in the 488 nm and 514.5 nm lines.

The tube design of the argon laser is much more complicated than that of the He–Ne laser, principally because of the much higher energy required to pump the ionic levels and the need to dissipate the heat energy released. The current density can be increased by concentrating the discharge with a magnetic field applied along the axis of the tube (the ions spiral about the magnetic lines of force). This has the added advantage of reducing the number of ions which collide with and damage the walls of the tube. The tube is made of a refractory

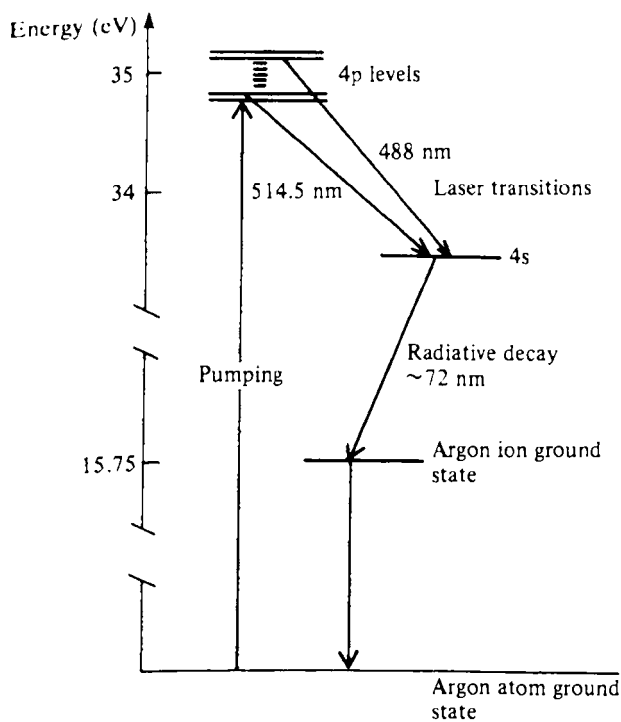


FIG. 5.40 Simplified energy level diagram for the argon ion laser. Ten or more laser lines are produced, but the two shown are by far the most intense.

material such as graphite or beryllium oxide. To dissipate the large amount of heat generated, most ion laser tubes are water cooled and often include a series of metal disks to act as heat exchangers, as shown in Fig. 5.41. Holes in the centre of these disks define the active region of the laser. Because of the high current involved, the cathode must be an excellent electron emitter and a getter is often incorporated to 'clean up' any impurities which might otherwise poison the cathode.

Again, the discharge is initiated by a high voltage pulse and then maintained by a d.c. voltage of about 200 V. During operation, the positive ions tend to collect at the cathode and may eventually cause the discharge to be extinguished. To prevent this, a gas return path is provided between the cathode and anode to equalize the pressure. Pulsed ion lasers tend to be simpler and, with a low duty cycle, the heat generated is small enough to be dissipated by convective cooling.

To select any desired wavelength, a small prism is inserted into the cavity and the position of the end mirror is changed by rotating it to be normal to the path of the radiation with the desired wavelength. This ensures that radiation of this particular wavelength will be reflected to and fro, while that of other wavelengths will be lost from the cavity after only a few round trips.

Krypton ion lasers are becoming increasingly used as excitation sources for dye lasers (see below) and in physical and chemical investigations. They produce a wealth of spectral

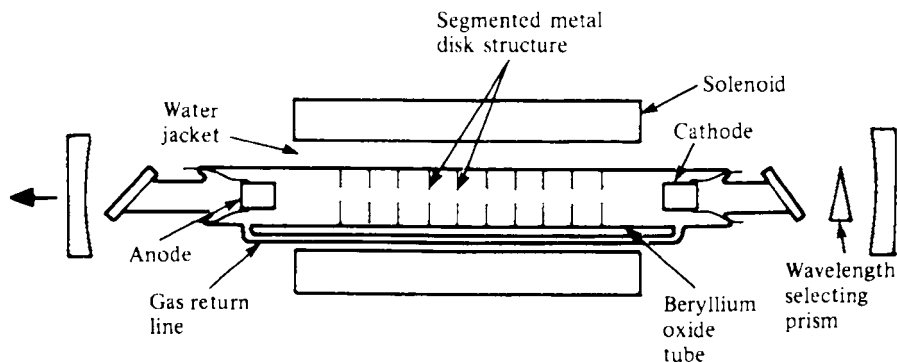


FIG. 5.41 Construction of a typical argon ion laser.

lines ranging across the entire visible spectrum from about 340 nm to 800 nm, with the most intense line at 647 nm.

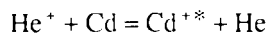
METAL VAPOUR ION LASERS

A group of ion lasers which is becoming increasingly important is based on discharges in ionized metal vapours such as gold, copper and cadmium; indeed many metallic elements in the periodic table can be made to lase in this way. We shall briefly describe the cadmium (Cd) laser, which has attracted considerable commercial interest because of its ability to generate short wavelength, CW emission at wavelengths of 441.6 nm and 325.0 nm. Other metal vapour ion lasers are described in ref. 5.23.

In many ways the operation of the Cd ion laser is very similar to that of the He–Ne laser in that the excitation process also involves helium. Excited states of the Cd^+ ions are achieved following collisions between cadmium atoms and energetic helium atoms in a discharge:



In this case, however, Cd^+ ions are also excited to even higher energies through collisions of cadmium atoms with energetic helium ions:



though the resulting laser emissions are not nearly as intense as those at 325 nm and 441.6 nm.

Metal ion lasers are technically more demanding than other gas lasers as not only does the metal have to be vaporized, but steps have to be taken to prevent the vapour from condensing on the laser windows. One way of doing this is to use a helium plasma to separate the windows from the main He–Cd discharge. It is also difficult to maintain a uniform distribution of vapour ions along the length of the discharge. The flow of vapour ions towards the cathode helps in this respect. A typical laser design is shown in Fig. 5.42. The cadmium vapour is generated at the anode end of the discharge tube in an oven held at a temperature of 220°C, and condensed at the cathode end. A second oven around the discharge tube maintains a constant temperature in the active region. The vapour pressure of cadmium is typically 0.002 torr, while that of helium is 3.5 torr; a discharge current of some 100 mA can then

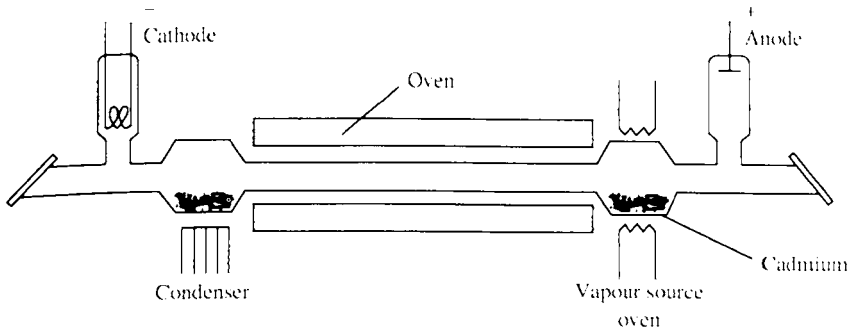


FIG. 5.42 Schematic diagram of an He-Cd⁺ vapour ion laser.

generate powers of some 200 mW at 441 nm, and 20 mW at a wavelength of 325 nm from a discharge tube 1.5 m long.

5.10.3.3 Molecular lasers – the carbon dioxide laser

The carbon dioxide laser is the most important molecular laser and indeed it is arguably the most important of all lasers from the standpoint of technological applications. In molecular lasers, the energy levels are provided by the quantization of the energy of vibration and rotation of the constituent gas molecules. The CO₂ molecule is basically an in-line arrangement of the two oxygen atoms and the carbon atom, which can undergo three fundamental modes of vibration as shown in Fig. 5.43. At any one time, the molecule can be vibrating in any linear combination of these fundamental modes. The modes of vibration are denoted by a set of three quantum numbers (v_1, v_2, v_3) which represent the amount of energy or number of energy quanta associated with each mode. The set (100), for example, means that a molecule in this state is vibrating in a pure symmetric mode with one quantum of vibrational energy; it has no energy associated with the asymmetric or bending modes.

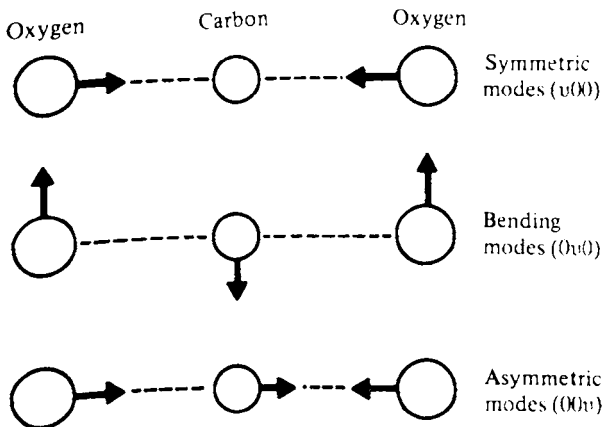


FIG. 5.43 Vibrational modes of the CO₂ molecule.

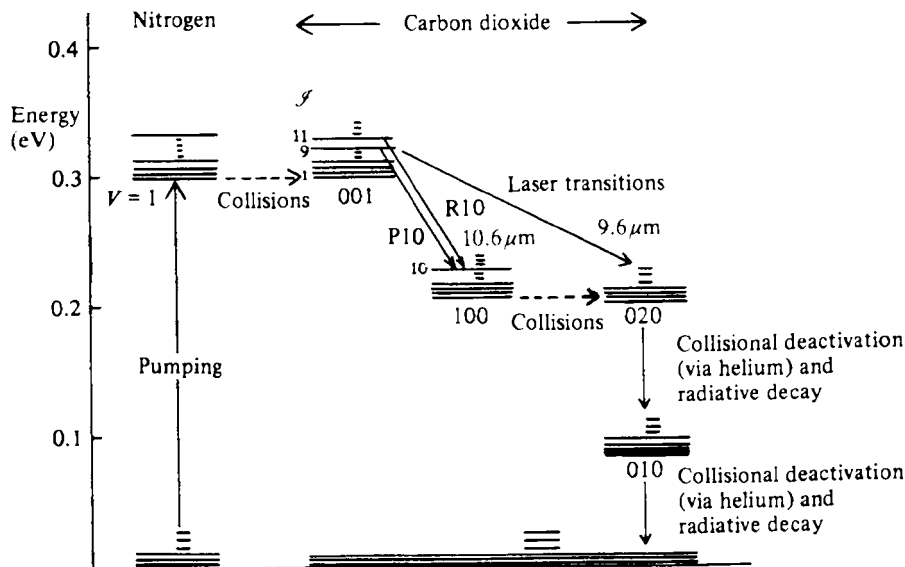


FIG. 5.44 Simplified energy level diagram for the CO₂ laser. Each vibrational level has many rotational levels associated with it, $\mathcal{J} = 1, 2, \dots$. The $10.6\ \mu\text{m}$ line is the strongest.

In addition to these vibrational modes, the molecule can also rotate, and thus it has closely spaced rotational energy levels associated with each vibrational energy level. The rotational levels are designated by an integer \mathcal{J} , the rotational quantum number. The energy separation between these molecular levels is small and the laser output is therefore in the infrared. The important parts of the CO₂ energy level arrangement are shown in Fig. 5.44, which also shows the ground state and first excited state of the vibrational modes of nitrogen.

Many CO₂ lasers contain a mixture of CO₂, nitrogen and helium in the ratio 1:4:5. Nitrogen plays a similar role to that of helium in the He–Ne laser. Excited nitrogen molecules transfer energy to the CO₂ molecules in resonant collisions, exciting them to the (001) levels. The (100) CO₂ levels have a lower energy and cannot be populated in this way, so that population inversion is created between the (001) and (100) levels giving stimulated emission at about $10.6\ \mu\text{m}$. The helium has a dual role. First, it increases the thermal conductivity to the walls of the tube, thereby decreasing the temperature and Doppler broadening, which in turn increases the gain. Secondly, it increases the laser efficiency by indirectly depleting the population of the (100) level, which is linked by resonant collisions to the (020) and (010) levels, the latter being depleted via collisions with the helium atoms.

While other gas lasers have efficiencies of 0.1% or less, the CO₂ laser may have an efficiency up to about 30%. This is essentially due to the ease with which electrons in the discharge can cause excitation and because of the strong coupling of the various levels involved. Because of this high efficiency, it is relatively easy to obtain CW outputs of 100 W for a laser 1 m long. Powers of this magnitude and greater (see below) mean that the mirrors should have very low absorption, while operation in the infrared also means that special materials

must be used for windows, mirrors and other laser components. Materials which have been used successfully include germanium, gallium arsenide, zinc sulfide, zinc selenide and various alkali halides, although these suffer from being relatively soft and hygroscopic. In some cases, a diffraction grating mounted on a piezoelectric transducer is used instead of the high reflectance mirror. The grating permits tuning of the laser output over the large range of distinct lines which Fig. 5.44 shows are possible. In the transition from the (001) to the (100) group of levels, the selection rule $\Delta J = \pm 1$ operates. Thus, for example, for $J = 10$ in the upper level, J can be 11 or 9 in the lower level. Then using the convention that J refers to the rotational quantum number of the lower level, the corresponding transitions give rise to the P10 ($\Delta J = +1$) and R10 ($\Delta J = -1$) branches respectively. The strongest lines are given by the P18, P20 and P22 transitions (ref. 5.24).

CO₂ LASER CONFIGURATIONS

SEALED TUBE LASERS

By a sealed tube laser we mean one similar to that adopted for He-Ne lasers where the discharge gases are completely sealed within the discharge tube. The problem with this design, as far as CO₂ lasers are concerned, is that during discharge the CO₂ molecules tend to break down into carbon monoxide molecules, CO. This occurs at quite a high rate and dramatically shortens the lifetime of the laser. One solution to this problem is to add hydrogen or water vapour to the gas mixture, which react with the CO and regenerate CO₂. Despite the presence of helium, which helps to conduct heat to the tube walls, gas cooling is another problem and generally the total power output is not more than about 100 W, while the lifetimes have been restricted to a few thousand hours. Sealed tube designs consequently are not very common, but they are used in conjunction with the so-called waveguide design, in which the inner dimensions of the tube are small (of the order of millimetres) and form a dielectric waveguide. Excellent beam quality with relatively large output powers can be obtained in this way from compact structures. Excitation of the gas is normally obtained by means of an intense radio-frequency field that can penetrate the dielectric material. Recently improvements in the discharge technology have enabled sealed tube lasers with CW outputs of some 500 W and lifetimes in excess of 20 000 hours to be fabricated (ref. 5.25).

GAS FLOW LASERS

Both degradation and cooling problems may be alleviated by allowing the gas mixture to flow through the laser tube. In the simplest designs both the gas flow and electrical discharge occur along the tube axis. If no attempt is made to recycle the gas, then a fresh mixture must be supplied continuously, though this is not a serious problem given that the gas pressures are relatively low. Vacuum pumps draw the gases through the discharge tube either to vent them to the atmosphere or for recycling.

The power output of the CO₂ laser increases approximately linearly with the length of the tube at a rate of about 60 W per metre. In attempts to obtain large output powers, lasers with tubes tens of metres in length have been built, which give powers of a few tens of kilowatts. Such tube lengths are generally impractical and other techniques are used to produce very high powers.

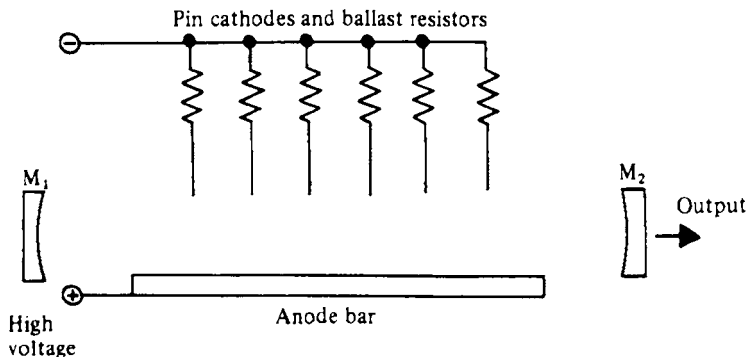


FIG. 5.45 Schematic diagram of the TEA CO_2 laser. The discharge is perpendicular to the axis of the laser cavity.

TRANSVERSELY EXCITED ATMOSPHERIC (TEA) LASERS

In the CO_2 laser the output power can be increased by increasing both the gas pressure and the tube length. The problem then is that it becomes increasingly difficult to create a discharge. At atmospheric pressure, for example, the breakdown voltage is 1.2 kV mm^{-1} and thus even for a laser 1 m long we would need an unacceptably high voltage. To overcome this, the discharge is struck *transversely* across the tube so that the discharge path length is now about a centimetre. This arrangement is referred to as transverse excitation at atmospheric pressure. The high voltage is applied to a set of electrodes placed along the tube as shown in Fig. 5.45. With this arrangement, peak powers in the gigawatt range can be obtained in very short pulses with a pulse repetition rate of about 20 Hz. To obtain a stable discharge and submicrosecond pulse rise times it is necessary to pre-ionize the gas before the main discharge occurs. This can be achieved by using either ultraviolet radiation or electron beams. Cool gas flowing through the lasing region further increases the population inversion and hence the power output.

GAS DYNAMIC LASERS

In gas dynamic lasers the population inversion is created through the application of thermodynamic principles rather than via a discharge. A nitrogen-carbon dioxide mixture is heated and compressed and then allowed to expand very rapidly into a low pressure region. During heating and compression the population of the energy states reaches the Boltzmann distribution appropriate to the higher temperature. At high temperatures most of the energy is stored in the vibrational modes of the nitrogen molecule. At lower temperatures, after expansion into the low pressure region, resonant collisions of the nitrogen molecules with the carbon dioxide molecules populate the (001) state of the CO_2 molecules and create a population inversion. With very active pumping of this type, CW output powers in excess of 100 kW have been achieved. Gas dynamic lasers suffer from the disadvantage of large size and the rocket-like roar as the gas expands.

OTHER MOLECULAR LASERS

As the number of molecular systems with energy levels which may be suitable for laser action is very large, it is not surprising that there are other molecular lasers and it is likely that many

more will be discovered. Two such molecular lasers emitting in the ultraviolet – visible region have been developed, namely *nitrogen* and *excimer lasers*.

The nitrogen laser differs markedly from the CO_2 laser. In the latter the transitions are between molecular rotational/vibrational energy levels, while in nitrogen the laser transitions are between electronic energy states and give rise to an output in the ultraviolet at 337 nm. A requirement for CW operation is that the upper level of the lasing transition should have a long lifetime while the lower level should be rapidly depopulated. In nitrogen, however, the converse is true: the upper level lifetime is exceedingly short (of the order of nanoseconds) while the lower level lifetime is of the order of microseconds. Hence population inversion and lasing action can only be maintained for a few nanoseconds. Very fast rise time pumping mechanisms similar to those used for TEA lasers must be used.

The gain in nitrogen is so large that it can be used as a simple amplifier; that is, in many applications it is not necessary to provide feedback (such high gain is termed *superradiant*) though a high reflecting back mirror is used to collect the amplified light into a single output beam. Commercial nitrogen lasers are capable of producing 100 kW peak power pulses. They are often used in photochemical investigations and for pumping other lasers, for example dye lasers (section 5.10.4).

In contrast to nitrogen, excimers provide a metastable excited state. An excimer (or *excited dimer*) refers to a molecule formed by the association of one excited atom (or molecule) with another atom (or molecule) which is in the ground state. If both constituents are in the ground state, then at interatomic distances characteristic of molecules they repel each other. Consequently, the excimer readily dissociates thereby effectively reducing the population of the lower lasing level and increasing the ease with which population inversion is established. If one or both of the constituents of the excimer are rare gas atoms, the excitation energy is extremely large and the metastable excimer state is an important system for storing energy.

Since 1972, a large number of lasers based on excimers have been developed; these cover the wavelength range from about 120 nm to 500 nm. Rare gas halide excimer lasers are especially efficient with XeF and KrF giving the highest efficiencies (some 10–15%).

Excimer lasers are usually pumped by an intense electron beam source or by a fast discharge. The electrons in the electron beam are accelerated until they have energies of 1 MeV and then transmitted to the laser chamber in pulses giving beam currents of about 100 kA. Electron beam generators with this sort of capability are rather large and it is more convenient to pump high pressure mixtures of the rare gas and halogen by collisional reactions in a fast pulsed discharge using a TEA configuration. Such lasers can generate some 100 W of average quasi-CW power in 30 ns pulses at rates of 1 kHz.

5.10.4 Liquid dye lasers

Liquids have useful advantages in relation to both solid and gas laser media. Solids are very difficult to prepare with the requisite degree of optical homogeneity and they may suffer permanent damage if overheated. Gases do not suffer from these difficulties but have a much smaller density of active atoms. Several different liquid lasers have been developed, but the most important is the dye laser (ref. 5.26). It has the advantage that it can be tuned over a

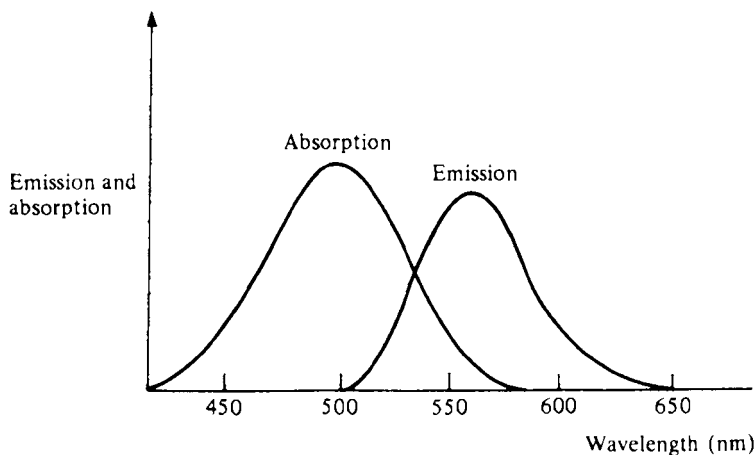


FIG. 5.46 Absorption and emission (fluorescence) spectra of a typical dye laser.

significant wavelength range. This is extremely useful in many applications such as spectroscopy and the study of chemical reactions.

The active medium is an organic dye dissolved in a solvent. When the dye is excited by short wavelength light it emits radiation at a longer wavelength, that is it *fluoresces*. The energy difference between the absorbed and emitted photons ultimately appears as heat – typical absorption and emission spectra are shown in Fig. 5.46. The broad fluorescence spectrum can be explained by the energy level diagram of a typical dye molecule. As Fig. 5.47 shows, the molecule has two groups of closely spaced electronic energy levels: the singlet

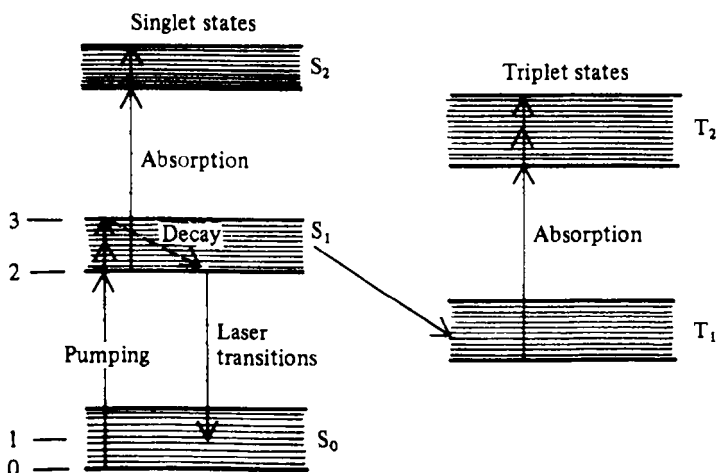


FIG. 5.47 Energy level scheme for a dye molecule. The laser transitions terminate above the lowest energy in the S_0 singlet state, so the laser is a four-level system. $S_1 \rightarrow T_1$ transitions lead to strong absorption at the laser wavelengths, thereby quenching laser action. $S_1 \rightarrow S_2$ transitions may also quench laser action in some dyes.

states (S_0 , S_1 and S_2) and the triplet states (T_1 and T_2). The singlet states occur when the total spin of the excited electrons in each molecule is zero (the value of $2S + 1$ is thus unity). The triplet states occur when the total spin is unity ($2S + 1 = 3$). Each electronic energy level is broadened into a near continuum of levels by the effects of the vibration and rotation of the dye molecule and also by the effects of the solvent molecules. Pumping results in the excitation of the molecule from the ground state S_0 to the first excited state S_1 . This is followed by very rapid non-radiative decay processes to the lower of the energy levels in S_1 . The laser transition is then from these levels to a level in S_0 . Since there are many such rotational/vibrational levels within S_0 and S_1 , there are many transitions resulting in an emission line which is very broad. As the termination of the laser transition in S_0 is at an energy much larger than kT above the bottom of S_0 , the dye laser is a four-level system and threshold is reached with a very small population inversion.

Although the triplet states are not directly involved in the laser action, they have a profound effect as there is a small probability of a transition $S_1 \rightarrow T_1$, even though this is forbidden by quantum mechanical selection rules. Since the transition $T_1 \rightarrow S_0$ is also forbidden, molecules 'pile up' in T_1 . The transition $T_1 \rightarrow T_2$ is allowed, however, and unfortunately the range of frequencies required for this transition is almost exactly the same as the laser transition frequencies. Thus, once a significant number of molecules have made the $S_1 \rightarrow T_1$ transition, $T_1 \rightarrow T_2$ absorption reduces the gain and may stop the lasing action. For this reason, most dye lasers operate in short pulses – shorter, in fact, than the time taken for T_1 to acquire a significant population, which is typically 1 μ s. For long pulse or CW operation, the population in T_1 will build up to equilibrium values, in which case the absorption is high and becomes the ultimate limitation on the efficiency of the laser.

Many dyes have been used as laser media, and Fig. 5.48 shows that, by tuning, laser wavelengths covering the whole of the visible spectrum can be obtained. The dye called rhodamine 6G with methanol as a solvent is one of the most successful, having an efficiency of about 20% and a broad tuning range (570–660 nm).

All dye lasers are optically pumped, the pumping source having a wavelength slightly

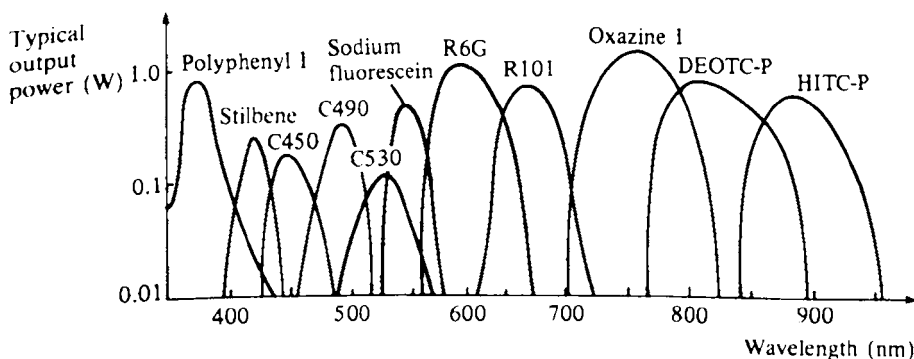


FIG. 5.48 Relative outputs of some common laser dyes pumped by ion lasers. Rhodamine 6G (R6G), for example, is here pumped by 5 W of power from all the argon lines. Coumarin dyes are labelled C; C490 is pumped by 2.3 W at 488 nm.

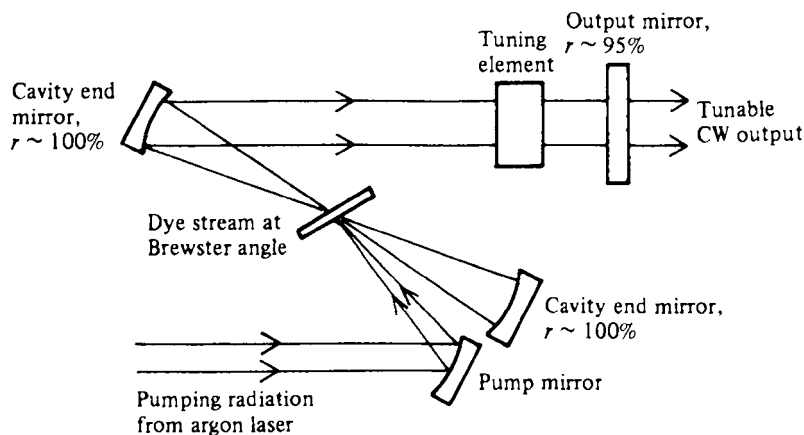


FIG. 5.49 Schematic diagram of a tunable CW laminar flow dye laser. The dye stream flow is perpendicular to the page.

less than that of the laser output. Commercial pumping methods include flashtubes, nitrogen lasers, solid state lasers and ion (A^+ or Kr^+) lasers. As the pumping radiation is in the visible or ultraviolet part of the spectrum, we must use harmonics of the output from neodymium lasers. The choice of the pump source depends on the absorption spectrum of the dye being used and the type of output desired. For CW output, the usual commercial pump source is an ion laser; all the other pumping sources operate in the pulsed mode and produce a pulsed output from the dye laser. As the population of T_1 builds up in about $1\ \mu s$, the flashtubes must discharge in about $1\ \mu s$ (in contrast to the millisecond time of tubes used with solid lasers). For this reason, specially strengthened flashtubes must be used.

To avoid absorption losses due to a build-up of molecules in triplet states and thereby obtain a CW output, the dye is passed in a liquid jet through the pumping radiation. The dye emerges from a specially shaped nozzle in the form of a smooth laminar sheet; the small signal gain is so high that this very small thickness of the active medium is sufficient to give laser action. The dye flows at a rate of about $10\ m\ s^{-1}$ so that the molecules spend a very short time in the cavity, which is less than the lifetime of the triplet states.

The laser output can be tuned using, for example, a prism, wedge filter or a diffraction grating, which can serve as a combined end mirror and dispersing element. A typical dye laser arrangement, which might generate an output of 1 W from a 3 W argon ion pump source, is shown in Fig. 5.49.

5.10.5 Parametric lasers

Although the principles of optical parametric oscillators (OPOs), which were described briefly in section 3.9, have been well understood for many years, it is only relatively recently that a range of efficient devices has become available commercially (ref. 5.27). The commercial viability of OPOs has resulted from two factors. First dramatic improvements have

been made in the quality of non-linear crystals such as BBO (beta-barium borate), KTP (potassium titanyl phosphate), KNB (potassium niobate) and LBO (lithium triborate). Secondly there have been comparable improvements in the beam quality, stability and reliability of appropriate pump lasers such as *Q*-switched Nd:YAG lasers (sections 5.10.1 and 6.4) and argon ion lasers (section 5.10.3.2). Consequently while conversion efficiencies of only 1% or so were achieved by Giordmaine and Miller, efficiencies of up to 60% are now common. Though only pulsed devices operating at pulse repetition rates of a few tens of hertz are currently readily available, OPOs operating at very much higher repetition rates, and in CW mode, have been demonstrated.

As mentioned in section 3.10 an OPO in its simplest form consists of the non-linear crystal placed between two mirrors forming an optical cavity, which resonates at either the signal or idler frequencies. As the power of the pump laser wave is increased a threshold pumping power will be reached, at which the parametric gain balances the losses, and the signal or idler waves will oscillate within the cavity to produce a coherent output. The practical importance of the OPO is its ability to convert the pump power into coherent waves which can be tuned continuously over large frequency ranges, in contrast to most of the other lasers described in this chapter.

Tuning of the idler and signal waves can be achieved in a number of ways as the refractive indices of the non-linear crystals depend on frequency, crystal orientation (if one or other of the waves is an extraordinary one), temperature and electric field. Giordmaine and Miller used temperature variation of the lithium niobate crystal to control the frequency of the oscillations as indicated in Fig. 5.50. A temperature change of about 11°C produces output frequencies in the range 3.1×10^{14} to 2.6×10^{14} Hz (corresponding to a wavelength range of 968–1154 nm).

Alternatively if we remember that in general $k = 2\pi/\lambda = \omega n/c$, we can write eq. (3.29) in

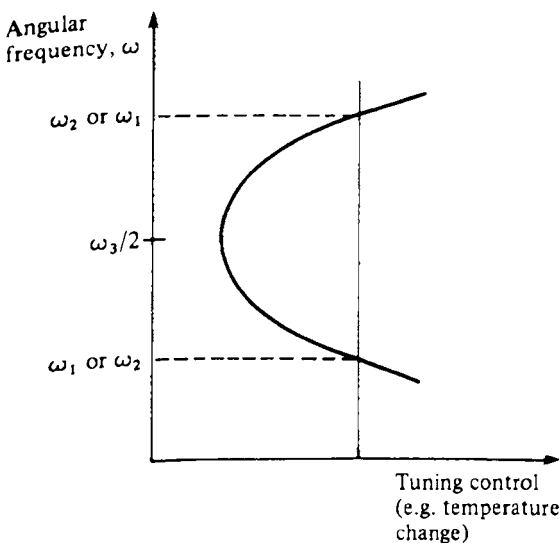


FIG. 5.50 Schematic tuning curve for a parametric oscillator.

the form

$$\omega_3 n_3 = \omega_1 n_1 + \omega_2 n_2 \quad (5.43)$$

Hence if the pump waves correspond to extraordinary waves, then changes in the beam direction of the pump relative to the crystal's optic axis will change n_3 , and hence ω_1 and ω_2 (the signal and idler waves are usually chosen to be ordinary waves so that $n_1 = n_2 = n_0$). A change in angle of some 8° can give rise to a tuning range extending from about 400 nm to 700 nm.

Commercially available OPOs cover the spectral range from about 400 nm to $4.0\ \mu\text{m}$ though oscillations from 300 nm to $18\ \mu\text{m}$ have been demonstrated. Although there are many actual and potential applications for a continuously tunable laser source, such as photochemistry, including the study of ultrafast chemical reactions (refs 6.4d and 6.8), high resolution spectroscopy and biomedicine, considerable improvements in the beam quality are required, and the linewidths need to be reduced. In addition, present average power levels are below those of other tunable laser sources, and parametric processes are difficult to exploit under CW conditions because of the low gain available when the non-linear crystal is pumped with CW radiation.

5.10.6 The free electron laser

We conclude this chapter with a brief description of a laser which has enormous potential in terms of its possible applications; these arise from its tunability over a wide wavelength range (from the extreme ultraviolet to millimetre wavelengths), with high power and high efficiency. Although the operating principles have been verified, the device is still at the developmental stage, where the main aims are to reduce its size and cost.

The operation of the free electron laser (FEL) are quite different from any of the lasers considered so far (ref. 5.28). The basic source of energy is a beam of relativistic electrons (i.e. electrons travelling at a speed very close to that of light), which is generated by an electron gun (which is, in fact, a particle accelerator). Under certain circumstances the electrons can be induced to give up some of their energy to a beam of photons travelling in the same direction, thereby amplifying the beam. The photon beam is from another laser, such as a CO_2 pulsed laser.

Now as the electric field of the light beam is at right angles to its direction of travel, the photons cannot receive any energy from the electrons unless they too have a component of velocity perpendicular to the direction of travel. Hence the electron beam is given an oscillatory path by passing it through a so-called 'wiggler' magnet, which consists of a row of alternating NS/SN magnets as shown in Fig. 5.51. This generates a spatially periodic magnetic field so that the electrons undergo the transverse oscillations indicated in the diagram. As a result of these oscillations the electrons emit monochromatic radiation in the forward direction at a wavelength which depends on their velocity, the magnet spacing and the value of the magnetic field. In fact the wavelength peaks at a value given by

$$\lambda = \frac{\lambda_w}{2\gamma^2} \quad (5.44)$$

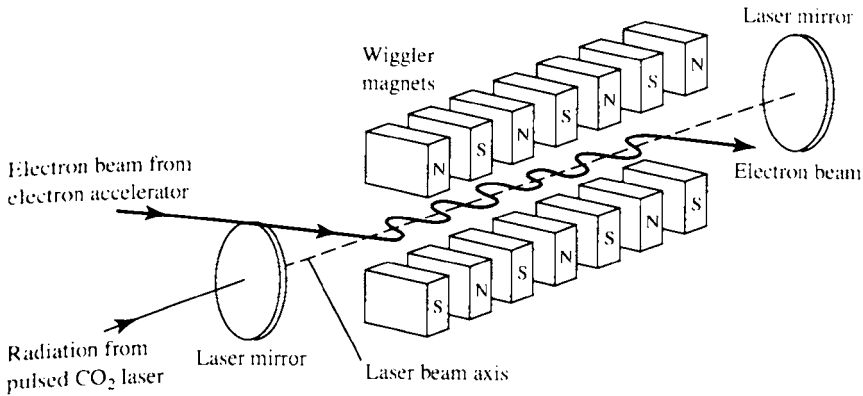


FIG. 5.51 Schematic diagram of the free electron laser.

where λ_w is the spatial period of the wiggler, and γ is the ratio of the total electron energy to its rest energy, which is 0.511 MeV (ref. 5.28). Such radiation is termed *magnetic*. The bandwidth, $\Delta\lambda$, of the radiation is given by

$$\frac{\Delta\lambda}{\lambda} = \frac{1}{2N} \quad (5.45)$$

where N is the number of wiggler periods.

Although one might expect that the radiation from the electron would now simply interact with the original laser beam and amplify it, this does not occur as readily as might at first appear. If the electron beam is 'tuned' to emit at exactly the same wavelength as the original light beam then in fact the energy exchanges between the two beams are equal and opposite.

One way of solving this problem is to inject the electrons at an energy slightly higher than that corresponding to the resonant wavelength (or 'resonant energy'); the electrons then tend to be 'pulled back' towards the resonant condition. Obviously only a small amount of energy, that is the difference between the initial and final electron energies, can be transferred in this way. Alternatively the wiggler period may be varied so that the final resonant energy is less than that at the start. The electrons then tend to 'home-in' to the resonant transfer condition and in so doing give up energy to the light beam. To make the FEL more efficient requires a careful reuse of the electron energy once the electron beam has transversed the interaction region; although various approaches have been suggested to accomplish this, none is yet very successful.

5.11

Conclusion

We have by no means covered the entire range of lasers currently available, nor fully discussed the various modifications and refinements of the lasers which we have described in

this chapter. We believe, however, that the basic laser physics covered, together with the descriptions of some typical lasers from the various 'classes' of laser, will enable the reader to understand the mode of operation of other lasers which might be encountered, or which undoubtedly will be developed in the future.

NOTES

1. In general, the probability of occupation of all the available states or energy levels is not the same. The states have a probability g_i of occupation (g is often called the *degeneracy*). Frequently, the ratio g_1/g_2 in eq. (5.5), for example, is of the order of unity and hence is often omitted.
2. The unit of energy most frequently used in this text in the energy level diagrams is the electron volt (eV). Many texts, however, use energy units expressed as cm^{-1} , a term often used by spectroscopists; $1 \text{ eV} \approx 8065 \text{ cm}^{-1}$.

PROBLEMS

- 5.1 At what temperature are the rates of spontaneous and stimulated emission equal (take $\lambda_0 = 500 \text{ nm}$)? At what wavelength are they equal at room temperature ($T = 300 \text{ K}$)?
- 5.2 If 1% of the light incident into a medium is absorbed per millimetre, what fraction is transmitted if the medium is 0.1 m long? Calculate the absorption coefficient α .
- 5.3 If the irradiance of light doubles after passing once through a laser amplifier 0.5 m long, calculate the small signal gain coefficient k assuming no losses. If the increase in irradiance were only 5%, what would k be?
- 5.4 An atom has two energy levels with a transition wavelength of 694.3 nm. Assuming that all the atoms in an assembly are in one or other of these levels, calculate the percentage of the atoms in the upper level at room temperature ($T = 300 \text{ K}$) and at $T = 500 \text{ K}$.
- 5.5 Calculate the degree of population inversion required to give a small signal gain coefficient for a CO_2 laser ($\lambda = 10.6 \mu\text{m}$) of 0.5 m^{-1} . Take the Einstein A coefficient for the upper laser level to be 200 s^{-1} .
- 5.6 Calculate the Doppler-broadened linewidth for the CO_2 laser transition ($\lambda_0 = 10.6 \mu\text{m}$) and the He-Ne laser transition ($\lambda_0 = 632.8 \text{ nm}$) assuming a gas discharge temperature of about 400 K. Take the relative atomic masses of carbon, oxygen and neon to be 12, 16 and 20.2 respectively.
- 5.7 Calculate the mirror reflectances required to sustain laser oscillations in a laser which is 0.1 m long, given that the small signal gain coefficient is 1 m^{-1} (assume the mirrors have the same values of reflectance).
- 5.8 Calculate the threshold small signal gain coefficient for ruby given the following data: threshold population inversion $5 \times 10^{22} \text{ m}^{-3}$, refractive index 1.5, linewidth $2 \times 10^{11} \text{ Hz}$, Einstein A coefficient 300 s^{-1} and wavelength 694.3 nm.

- 5.9 What is the mode number nearest the line centre of the 632.8 nm transition of the He–Ne laser; what is the mode separation; how many longitudinal modes could possibly oscillate given that the width of the gain curve is 1.5×10^9 Hz? Take the mirror separation to be 0.5 m.
- 5.10 By considering the rate equations for a three-level laser, derive an expression for the threshold population inversion in terms of the laser parameters. Hence compare the pump power required to bring a three-level and a four-level laser to threshold.
- 5.11 Calculate the threshold pumping power for an Nd:glass laser, given that the critical population inversion is $9 \times 10^{21} \text{ m}^{-3}$, the spontaneous lifetime is 300 μs and that the upper laser level is at an energy of 1.4 eV.
- 5.12 Assuming that the exciting lamp is 10% efficient, that 10% of the light produced actually falls on the crystal, which has a diameter of 2 mm and is 0.1 m long, and that 5% of the exciting light energy falls within useful absorption bands, estimate the power to be supplied to the lamp used to pump the laser in Problem 5.11.
- 5.13 Estimate the efficiency of a GaAs laser operating well above threshold, given that $n = 3.6$ and that the length of the laser cavity is 200 μm . Take the loss coefficient to be 800 m^{-1} and the internal quantum efficiency to be 0.8.

REFERENCES

- 5.1 T. H. Maiman, 'Stimulated optical radiation in ruby masers', *Nature*, **187**, 493, 1960.
- 5.2 (a) J. Hawkes and I. Latimer, *Lasers: Theory and Practice*, Prentice Hall International, Hemel Hempstead, 1995.
(b) J. Wilson and J. F. B. Hawkes, *Lasers: Principles and Applications*, Prentice Hall International, Hemel Hempstead, 1987.
(c) D. Wood, *Optoelectronic Semiconductor Devices*, Prentice Hall International, Hemel Hempstead, 1994.
(d) J. T. Verdeyen, *Laser Electronics* (2nd edn), Prentice Hall, Englewood Cliffs, NJ, 1989.
- 5.3 A. Einstein, 'Zur Quantentheorie der Strahlung', *Phys. Z.*, **18**, 121, 1917.
- 5.4 N. Bloembergen, 'Proposal for a new type of solid state maser', *Phys. Rev.*, **105**, 762, 1957.
(a) A. E. Siegman, *Introduction to Masers and Lasers*, McGraw-Hill, New York, 1971, Chapter 8.
(b) T. Li and H. Kogelnik, 'Resonator stability curves', *Appl. Opt.*, **5**, 1550, 1966.
- 5.6 (a) A. Yariv, *Optical Electronics* (4th edn), Holt, Rinehart & Winston, New York, 1991, Chapter 5.
(b) R. W. Ditchburn, *Light* (2nd edn), Blackie, Glasgow, 1962, pp. 91–7.
- 5.7 *Ibid.*, Chapter 4, pp. 85–91 and 106–17.
- 5.8 (a) A. E. Siegman, *Introduction to Masers and Lasers*, McGraw-Hill, New York, 1971.
(b) A. E. Siegman, 'Resonant modes in a maser interferometer', *Bell Syst. Tech. J.*, **40**, 453, 1961.

- 5.9 G. R. Fowles, *Introduction to Modern Optics* (2nd edn), Holt, Rinehart & Winston, New York, 1975, Chapter 4.
- 5.10 J. T. Verdeyen, *op. cit.*, Section 5.1.
- 5.11 (a) H. Semat and J. R. Albrit, *Introduction to Atomic and Nuclear Physics* (5th edn), Holt, Rinehart & Winston, New York, 1972, pp. 259–64.
(b) D. Eastham, *Atomic Physics of Lasers*, Taylor & Francis, London and Philadelphia, 1986.
- 5.12 J. S. Griffith, *The Theory of Transition-metal Ions*, Cambridge University Press, Cambridge, 1964, Chapter 9.
- 5.13 (a) G. P. A. Malcolm and A. I. Ferguson, 'Diode pumped solid state lasers', *Contemp. Phys.*, **32**, 305, 1991.
(b) J. Fitzpatrick, 'Laser diode arrays: pump up the power', *Photonics Spectra*, Nov., 105, 1995.
- 5.14 (a) R. J. Mears and L. Reekie, 'Neodymium-doped silica single-mode fibre laser', *Electron. Lett.*, **21**, 738, 1985.
(b) D. Hanna and A. Tropper, 'Fiber lasers offer higher powers and shorter wavelengths', *Laser Focus World*, May, 123, 1995.
(c) J. D. Minelly, 'Fibre lasers launch into medicine, aerospace and materials processing', *Photonics Spectra*, June, 129, 1996.
- 5.15 J. Hecht, 'Tunability makes vibronic lasers versatile tools', *Laser Focus World*, Oct., 93, 1992.
- 5.16 (a) H. C. Casey and M. B. Panish, *Heterojunction Lasers*, Academic Press, New York, 1978.
(b) H. Kressel and J. K. Butler, *Semiconductor Lasers and Heterojunction LEDs*, Academic Press, New York, 1977.
(c) J. Gower, *Optical Communication Systems* (2nd edn), Prentice Hall International, Hemel Hempstead, 1995.
(d) J. M. Senior, *Optical Fibre Communications: Principles and Practice* (2nd edn), Prentice Hall International, Hemel Hempstead, 1992.
(e) D. Wood, *op. cit.*, Chapter 4.
- 5.17 *Ibid.*
- 5.18 (a) *Ibid.*, Chapters 8 and 9.
(b) H. C. Casey and M. B. Panish, *op. cit.*, Part B, Chapters 5–7.
(c) H. Kressel and J. K. Butler, *op. cit.*, Chapter 9.
- 5.19 (a) P. S. Zory, *Quantum Well Lasers*, Academic Press, San Diego, 1993.
(b) N. Tessler *et al.*, *IEEE J. Quantum Electron.*, **28**(10), 2242, 1993.
- 5.20 (a) D. E. Ackley, 'Phase locked injection laser arrays with non-uniform stripe spacing', *Electron. Lett.*, **20**, 695, 1984.
(b) G. R. Olbright, 'VCSELs could revolutionise optical communications', *Photonics Spectra*, Feb., 1995.
(c) J. L. Jewell *et al.*, 'Vertical-cavity surface-emitting lasers: design, growth, fabrication, characterisation', *IEEE J. Quantum Electron.*, **27**(6), 133, 1991.
(d) K. Iga, 'Surface emitting lasers', *Opt. Quantum Electron.*, **24**, 597, 1992.

- (e) K. Uomi *et al.*, 'Low threshold room temperature pulsed operation of 1.5 μm multi-quantum well active layer', *IEEE Photonics Technol. Lett.*, **6**(3), 317, 1994.
- (f) L. A. Coldren and S. W. Corzine, *Diode lasers and photonic integrated circuits*, John Wiley, New York, 1995.
- 5.21 (a) S. Nakamura *et al.*, 'Gallium nitride based injection laser operating at 417 nm', *Jpn. J. Appl. Phys.*, **35**, L74, 1995.
- (b) M. A. Haas *et al.*, 'Blue-green laser diodes', *Appl. Phys. Lett.*, **59**, 1272, 1991.
- 5.22 D. Wood, *op. cit.*, pp. 104–7.
- 5.23 W. B. Bridges, *Methods of Experimental Physics*, Vol. 15, *Quantum Electronics*, Part 2, Academic Press, New York, 1979.
- 5.24 W. J. Wittenman, *The CO₂ laser*, Springer Series in Optical Sciences, Vol. 53, Springer-Verlag, Berlin, 1986.
- 5.25 K. Bondelic, 'Sealed carbon dioxide lasers achieve new power levels', *Laser Focus World*, Aug., 95, 1996.
- 5.26 F. J. Duarte and L. W. Hillman, *Dye Laser Principles*, Academic Press, New York, 1990.
- 5.27 (a) U. Stamm, 'OPOs advance in Europe, but challenges remain', *Photonics Spectra*, Mar., 110, 1995.
- (b) S. Butcher, 'Optical parametric oscillators open new doors to researcher', *Photonics Spectra*, May, 133, 1994.
- 5.28 (a) C. A. Brau, 'The free electron laser: an introduction', *Laser Focus*, **17**, 48, 1981.
- (b) T. Riordan, 'The free electron laser', *Photonics Spectra*, July, 40, 1983.
- (c) H. P. Freund and T. M. Antonse, *Principles of free electron lasers*, Chapman and Hall, London, 1992.

Lasers II

In the previous chapter, we saw that the output of lasers does not always consist of a beam of very coherent, almost monochromatic radiation. The output, for example, may be continuous or in the form of mutually incoherent spikes within a pulse and consist of several longitudinal and transverse modes of slightly different wavelength. In considering the applications of lasers we often find that such characteristics are quite acceptable, but equally it is often desirable to modify the laser output to suit a particular application. Some modifications to the output can be achieved quite simply: for instance, we can select one of the many wavelengths produced by the argon ion laser by introducing a prism or grating into the optical cavity. The prism or grating disperses the light so that after transmission only one wavelength falls normally onto the end mirror and is reflected back into the cavity. Other modifications of the output, though often quite readily achieved in practice, require a clear understanding of the concepts of modes, population inversion, threshold gain and the like. Before discussing a small selection of the many and varied applications of lasers, we first consider some of the ways in which the laser output may be modified to facilitate these and countless other applications.

6.1 Single mode operation

In many applications including chemical and physical investigations it is desirable to have the greatest possible spectral purity. We can achieve this by operating a CW laser in a single longitudinal, single transverse mode. Since an inhomogeneously broadened laser (see section 5.9 and below) can support several longitudinal and transverse modes simultaneously, single mode operation can be achieved only by arranging for one mode to have a higher gain than all the others. We can ensure that the cavity will support a single transverse mode only, the $TEM_{(0)}$ mode, by placing an aperture within the cavity. As the higher order TEM modes spread out further than the $TEM_{(0)}$ mode, an aperture of suitable diameter will transmit the $TEM_{(0)}$ mode while eliminating the others. All but one of the longitudinal modes can then be rejected by reducing the length L of the laser cavity until the frequency separation between adjacent modes, that is $\delta\nu = c/2L$ (see section 5.9), is greater than the linewidth of the laser transition. Figure 5.11 then shows that the single mode which falls within the transition linewidth is the only one that can oscillate.

The disadvantage of this system is that the active length of the laser cavity may become so small as to limit the power output severely. This may be overcome using other techniques involving, for example, a Fabry–Perot resonator either inside or outside the laser cavity,

third-mirror techniques or absorbers within the cavity (ref. 6.1). It should be stressed that to maintain the wavelength of the single mode output at a constant value, we must stabilize the cavity dimensions by rigid construction and temperature control or by introducing compensating systems. If this is not done, L will change and the frequency and power of the laser will change as a consequence. In passing, we note that it is possible to stabilize the operating frequency of a laser to better than 1 MHz, or about 1 part in 10^9 . Ways of doing this are described in the next section.

6.2 Frequency stabilization

In lasers with homogeneously broadened transitions, an increase in pumping cannot increase the population inversion beyond the threshold value where the gain per pass equals the losses. This is because the spectral lineshape function $g(\nu)$ describes the response of each individual atom, all of which are considered to behave identically. Thus, as the pumping is increased from below the threshold value the laser will begin to oscillate at the centre frequency ν_0 . The gain at other frequencies will remain below threshold, however, so that an ideal homogeneously broadened laser will oscillate only at a single frequency.

In inhomogeneously broadened lasers, on the other hand, where individual atoms are considered to behave differently from one another, the population inversion and gain profile can increase above the threshold values at frequencies other than ν_0 . The gain at ν_0 , however, remains clamped at the threshold value owing to gain saturation (section 5.6). Further pumping may increase the gain at other frequencies until oscillations commence at those frequencies also. This results in decreases in both the population inversion and the gain to their threshold values. The gain curve therefore acquires depressions or 'holes' in it at these oscillating frequencies – this is referred to as *hole burning*. The gain curves for homogeneous and inhomogeneous atomic systems are illustrated in Fig. 6.1, where the curves labelled A, B and C correspond to pumping levels respectively below threshold, at threshold and above threshold.

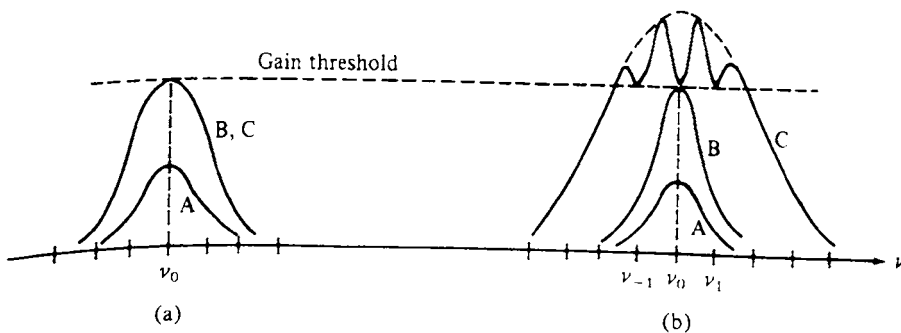


FIG. 6.1 Gain curves for (a) a homogeneously broadened atomic system and (b) an inhomogeneously broadened atomic system (A, below threshold; B, at threshold; C, above threshold). In (b), different groups of atoms respond to the stimulating radiation at different cavity mode frequencies. The gain is saturated at each of these frequencies independently, creating 'holes' in the gain curve.

Let us now consider gas lasers which are usually inhomogeneously broadened owing primarily to the Doppler effect as explained in section 5.7. We suppose that a single mode is oscillating at frequency ν_m which is greater than the natural emission frequency of the atoms, ν_0 . The oscillation, being a standing wave within the cavity, consists of two sets of waves travelling in opposite directions, say the positive and negative x directions respectively. Both of these waves have the frequency ν_m .

The interaction of the waves travelling in the positive x direction with the atoms in the laser medium will be greatest for those atoms that have a velocity component in the x direction of $+v_x$ such that

$$\nu_m = \nu_0 \left(1 + \frac{v_x}{c} \right) \quad (6.1)$$

For this group of atoms, the apparent frequency of the waves is ν_m and the atoms are stimulated to emit. The argument also holds for a second group of atoms and waves moving in the negative x direction. There are therefore two groups of atoms whose stimulated emission contributes to the laser output intensity; the population inversion is reduced for these atoms and gain saturation occurs.

We have plotted the population inversion N as a function of the x component of velocity in Fig. 6.2(a) where we see that stimulated emission produces a saturation in the excited state atomic velocity distribution similar to the hole burning in the gain curve. Two 'holes' are burned'; these are symmetrically placed about $v_x = 0$ and correspond to atoms with velocities of plus and minus v_x .

Suppose that now the frequency of the oscillating mode is changed until it equals the peak

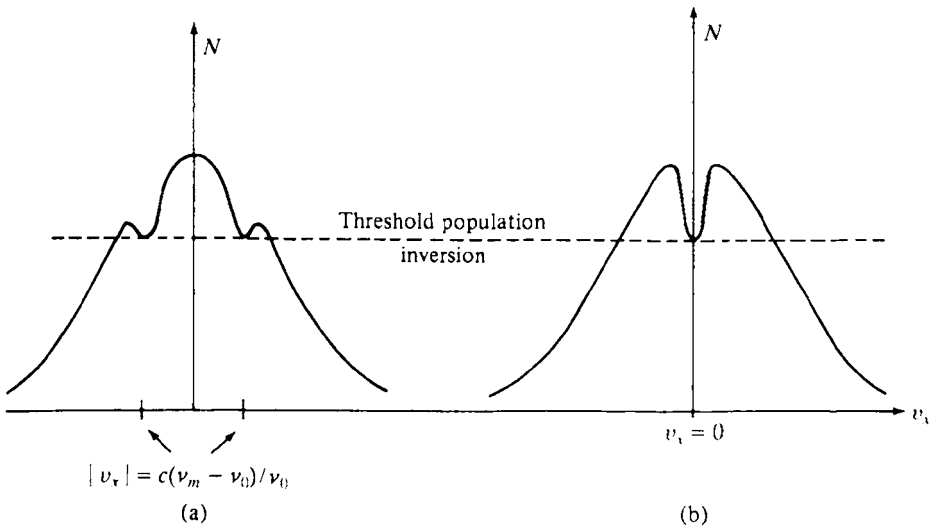


FIG. 6.2 Distribution of the population inversion N as a function of the x component of velocity of the atoms v_x . In (a) the cavity frequency ν_m is different from the natural emission frequency of the atoms ν_0 and two 'holes' are created in the distribution. In (b) the cavity frequency equals the atomic resonant frequency, i.e. $\nu_m = \nu_0$ and only one hole is created, corresponding to $v_x = 0$.

frequency of the laser line, that is $\nu_m = \nu_0$. This may be accomplished, for example, by varying the temperature to change the cavity length slightly. Under these circumstances, only a single group of atoms can contribute to the lasing process, namely those with zero x component of velocity, and there is a single 'hole' in the population inversion-velocity curve as shown in Fig. 6.2(b). When this happens, the laser output power drops as the available inverted population is smaller than before. A plot of output power as a function of frequency ν_m as in Fig. 6.3 then shows a dip, the *Lamb dip*, at the centre frequency $\nu_m = \nu_0$.

The increase in power resulting from any slight deviation from the centre of the laser line can be used, as the basis of a feedback system, to stabilize the frequency of the laser at the line centre by minimizing the output. Such techniques enable the frequency to be stabilized to better than 1 part in 10^9 . The long coherence length makes the output of lasers stabilized in this way useful in applications such as long path difference interference measurements (section 6.6.1).

Another technique for stabilizing gas lasers relies on the fact that the gain profile is symmetrical about its midpoint. Consider the situation when only two modes are operating with equal irradiances. They must be equally spaced in frequency either side of the gain profile maximum. Any drift in the mode frequency will cause one mode to increase in irradiance and the other to decrease. Thus, if we are able to monitor the two-mode irradiance and then use the difference to operate a feedback loop that controls the cavity length, we should be able to stabilize the operating frequency. At first sight, the measurement of the mode irradiances might seem difficult. Fortunately it is usually found that adjacent cavity modes are plane polarized with their planes of polarization at right angles to each other. Thus we need only split the laser output into two, insert suitably orientated pieces of polaroid into the two beams, and then allow them to fall onto two detectors. The output of the detectors is then proportional to the irradiances of the modes. A simple way of applying feedback is to let any difference in output of the detectors modulate the current passing through a heating coil wrapped round the laser tube. Any changes in relative mode irradiance will then alter the

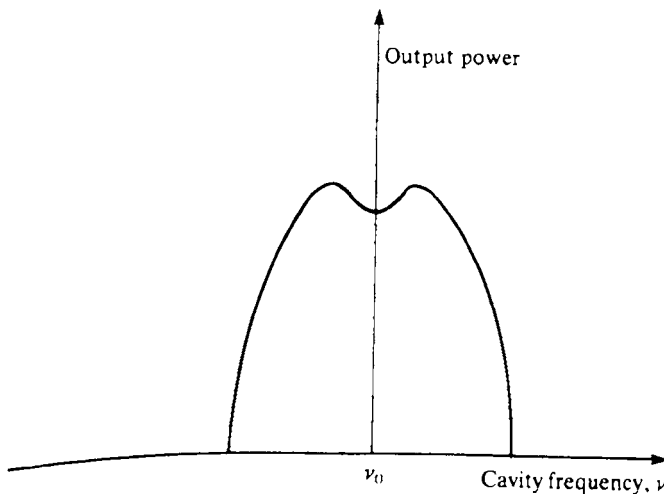


FIG. 6.3 Output power as a function of frequency of a single mode laser. The minimum at the centre of the curve ($\nu = \nu_0$) is known as the *Lamb dip*.

tube temperature and hence the effective length of the cavity. This in turn will vary the mode frequencies (ref. 6.2).

As might be expected, the techniques used to frequency-stabilize semiconductor lasers are somewhat different from those used in gas lasers. Obviously, the first step is to ensure that only a single mode can oscillate. Obtaining a single transverse mode is relatively easy: the gain region may be narrowed using the techniques discussed in section 5.10.2.3. The gain curve is relatively wide, however, and several longitudinal modes are usually present. Even if we ensure that only one of these can oscillate, there are problems with temperature stability. These arise because the position of the gain profile depends on the bandgap, and this in turn is temperature dependent. Thus, a change in temperature could cause the position of the gain profile to alter sufficiently to cause the laser to 'hop' to another longitudinal mode more favourably placed with regard to the gain profile.

One way to improve this situation is to provide some wavelength-sensitive feedback that is relatively insensitive to temperature. A very effective way of implementing this is to use a 'distributed feedback' structure (ref. 6.3). The way in which this operates may be appreciated by reference to Fig. 6.4. This shows a laser diode structure with a region at either end containing corrugations. These act as frequency-selective mirrors. To see this, we may consider two beams emerging from the pumped region at a small angle to the axis and being reflected from different parts of the corrugations as shown in Fig. 6.5. There will be constructive interference between the two beams when $2D = m\lambda_0/n$, where D is the 'wavelength' of the corrugations, λ_0 the vacuum wavelength, n the refractive index of the laser medium, and m an integer. The corrugated structure will only act as a mirror for a particular mode if its wavelength satisfies the above equation. Thus, by choosing an appropriate value of D we can ensure that only one particular longitudinal mode can oscillate. The formal treatment is rather complex and the interested reader is referred to ref. 6.4.

Although the structure of Fig. 6.4, which is often referred to as a distributed Bragg reflector (DBR), undoubtedly works, it has the disadvantage that light enters the unpumped end regions, where it may suffer a considerable amount of absorption. For this reason a more

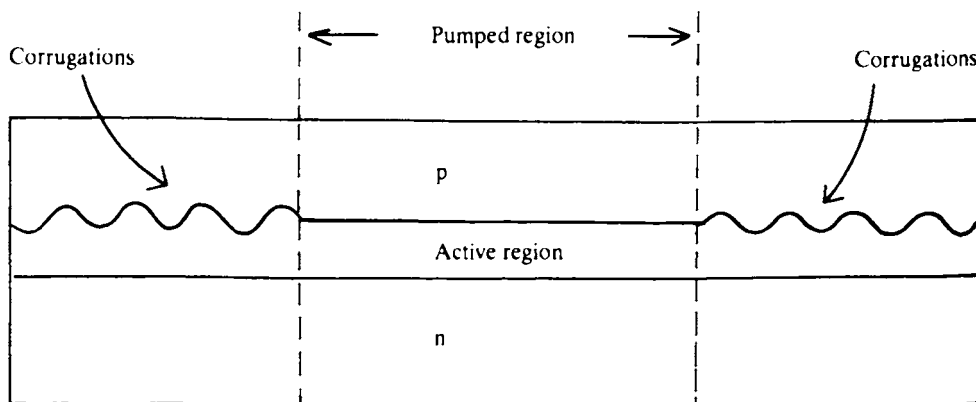


FIG. 6.4 Schematic diagram of a heterostructure laser incorporating regions at either end which have corrugations etched along one side. Such regions act as frequency-selective mirrors.

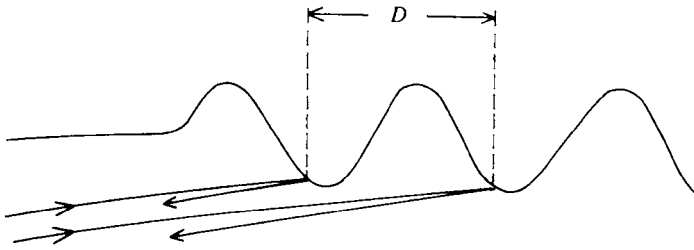


FIG. 6.5 Two beams of light being reflected from corresponding adjacent regions of the corrugation shown in Fig. 6.4. The two reflected beams will remain in phase provided that the path differences are multiples of the wavelength. Assuming that the beams make small angles with the laser axis, this path difference is very nearly equal to $2D$.

efficient structure is given by extending the corrugations throughout the pumped region, thus giving rise to the *distributed feedback* (DFB) laser shown in Fig. 6.6. Here, it will be noticed that the corrugated region is in a layer adjacent to the active layer, there being sufficient coupling between the two for the mirror action of the corrugations to be effective. This is because it is difficult to form the corrugations directly on the active layer without introducing defects into the active region that degrade the performance of the laser.

An alternative approach is to pass the light through an additional, external coupled cavity. In this way the only modes which can interfere constructively and oscillate are those which resonate in both cavities. This results in a much greater mode frequency spacing so that only one mode lies within the gain curve of the active medium.

Several designs have been used to achieve this including an external mirror (ref. 6.5a) and the so-called cleaved coupled cavity (C^3) arrangement (ref. 6.5b). In the latter case two laser cavities are formed by cleaving from the same original chip so that the end facets (mirrors) are perfectly parallel to each other. The two cavities are then mounted parallel to each other on a heat sink. The current to each can be controlled separately. In operation the current flow to one of the cavities is above threshold so that it acts as the laser, while that to the other is kept below threshold, but its magnitude can be used to alter the carrier density and hence the refractive index of the cavity. In this way the operating frequency can be tuned over a fairly wide wavelength range.

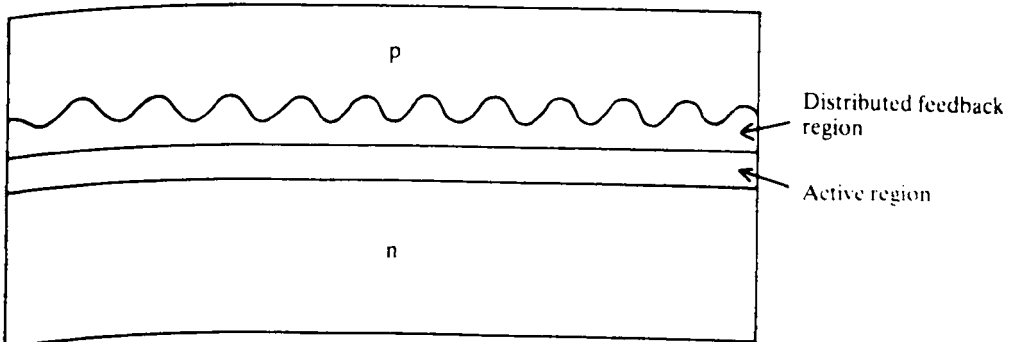


FIG. 6.6 Distributed feedback laser structure.

6.3

Mode locking

Mode locking is a technique for producing periodic, high power, short duration laser pulses. As we saw in the previous section, a typical inhomogeneously broadened laser cavity may support oscillations in many modes simultaneously. The output of such a laser as a function of time depends on the relative phases, frequencies and amplitudes of the modes. The total electric field as a function of time can be written as

$$\mathcal{E}(t) = \sum_{n=0}^{N-1} (\mathcal{E}_0)_n \exp[i(\omega_n t + \delta_n)] \quad (6.2)$$

where $(\mathcal{E}_0)_n$, ω_n and δ_n are the amplitude, angular frequency and phase of the n th mode. Usually these parameters are all time varying, so that the modes are incoherent and the total irradiance is simply the sum of the irradiances of the individual modes as we saw in section 1.2.2. Hence, for this situation, which is illustrated in Fig. 6.7(a),

$$I = N\mathcal{E}_0^2$$

where we have assumed for simplicity that all N modes have the same amplitude \mathcal{E}_0 . The irradiance may exhibit small fluctuations if two or three of the modes happen to be in phase at any given time.

Suppose that we now force the various modes to maintain the same relative phase δ to one another, that is we *mode lock* the laser such that $\delta_n = \delta$. The total irradiance must now be found by adding the individual electric fields rather than the irradiances. Using eq. (6.2), the resultant electric field can now be written as

$$\mathcal{E}(t) = \mathcal{E}_0 \exp(i\delta) \sum_{n=0}^{N-1} \exp(i\omega_n t) \quad (6.3)$$

For convenience, let us write the angular frequency ω_n as $\omega_n = \omega - n\delta\omega$, where ω is the angular frequency of the highest frequency mode and $\delta\omega$ is the angular frequency separation between modes, which from eq. (5.32) we can write as

$$\delta\omega = \pi c/L$$

Equation (6.3) for $\mathcal{E}(t)$ can then be rewritten as

$$\begin{aligned} \mathcal{E}(t) &= \mathcal{E}_0 \exp(i\delta) \sum_{n=0}^{N-1} \exp[i(\omega - n\delta\omega)t] \\ &= \mathcal{E}_0 \exp\left[i(\omega t + \delta) \sum_{n=0}^{N-1} \exp(-\pi i n c t / L)\right] \end{aligned}$$

or

$$\mathcal{E}(t) = \mathcal{E}_0 \exp[i(\omega t + \delta)] \{1 + \exp(-i\phi) + \exp(-2i\phi) + \dots + \exp[-(N-1)i\phi]\} \quad (6.4)$$

where $\phi = \pi c t / L$. The term in braces in eq. (6.4) is a geometric progression and we can thus

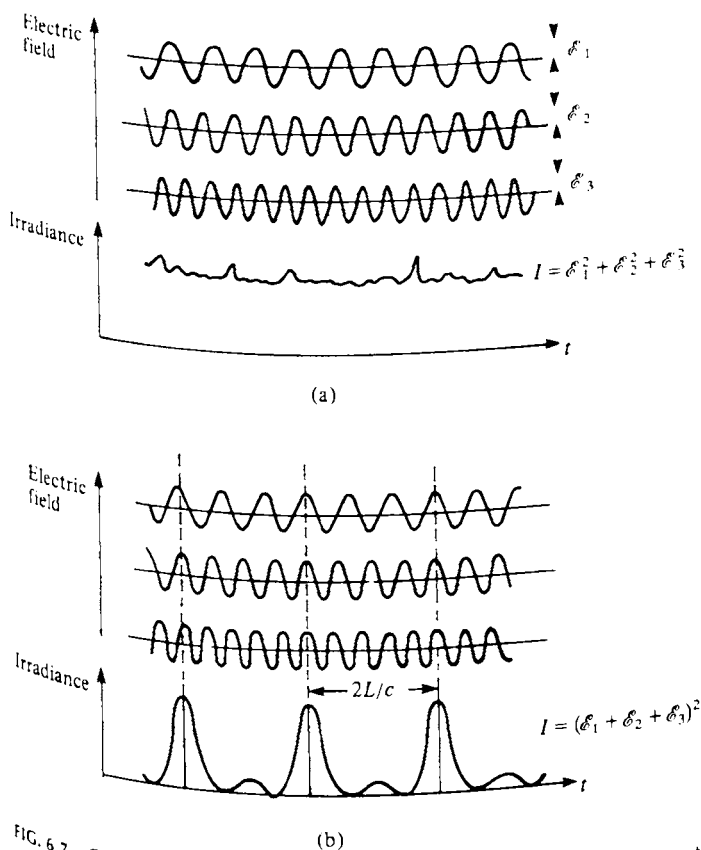


FIG. 6.7 Comparison of (a) non-mode-locked and (b) mode-locked laser outputs. In (a) the irradiance shows random fluctuations while in (b) the phase relationships of the modes are kept constant and the output consists of a series of narrow intense pulses of time spacing $2L/c$ and duration $(1/N)(2L/c)$.

write

$$\mathcal{E}(t) = \mathcal{E}_0 \exp[i(\omega t + \delta)] \frac{\sin(N\phi/2)}{\sin(\phi/2)}$$

The irradiance I is then $I = \mathcal{E}(t) \cdot \mathcal{E}^*(t)$ or

$$I(t) = \mathcal{E}_0^2 \frac{\sin^2(N\phi/2)}{\sin^2(\phi/2)} \quad (6.5)$$

The form of this equation (for $N = 3$) is illustrated in Fig. 6.7(b).

We see that the irradiance $I(t)$ is periodic ($\Delta\phi = 2\pi$) in the time interval $t = 2L/c$, which equals the round trip transit time for light within the cavity. The maximum value of the irradiance is $N^2 \mathcal{E}_0^2$. This occurs for values of $\phi = 0$ or $p\pi$, p being an integer, where the value of the function $\sin^2(N\phi/2)/\sin^2(\phi/2)$ equals N^2 .

Similarly, the irradiance has minimum values of zero when $N\phi/2 = p\pi$, p being an integer which is not zero, that is when $\phi = 2p\pi/N$ or $t = (1/N)(2L/c)p$. Thus the time duration of the maxima, which is the time taken for the irradiance to fall from its maximum value to an adjacent zero ($p = 1$), is $(1/N)(2L/c)$. We can see, therefore, that the output of a mode-locked laser consists of a sequence of short pulses, separated in time by $2L/c$, each of peak power equal to N times the average power (or N times the power of the same laser with the modes uncoupled). The ratio of the pulse spacing to the pulse width is approximately equal to the number of modes, that is $(2L/c)/[(2L/c)(1/N)] = N$. Thus, to obtain high power short duration pulses there should be a large number of modes; this requires a broad laser transition and a long laser cavity.

EXAMPLE 6.1 Mode-locked pulses

Let us compare the pulse separation and pulse duration in a mode-locked Nd:YAG laser where the fluorescent linewidth is 1.1×10^{11} Hz and the laser rod is 0.1 m long; we take the refractive index to be 1.8.

The mode separation $c/(2Ln) = 8.3 \times 10^8$ Hz; thus the number of modes oscillating is $(1.1 \times 10^{11})/(8.3 \times 10^8)$, that is about 132. The pulse separation is $2Ln/c \approx 1.24$ ns and the pulse duration is $(1/N)(2Ln/c) \approx 9$ ps.

The situation can be visualized by considering a short wave packet bouncing back and forth between the cavity mirrors; the pulses emitted by the laser appear each time the wave packet is partially transmitted by the output mirror as indicated in Fig. 6.8. This physical picture is particularly useful when describing the active mode-locking mechanism used with argon ion and Nd:glass lasers.

6.3.1 Active mode locking

Mode locking is achieved by forcing the longitudinal modes to maintain fixed phase relationships. This can be accomplished by modulating the loss (or gain) of the laser cavity at a frequency equal to the intermode frequency separation $\delta\nu = c/2L$ (or $\delta\omega = \pi c/L$). Let us imagine that the loss modulation is provided by a shutter placed near one of the mirrors. The shutter is closed (corresponding to very high losses) most of the time and is only opened

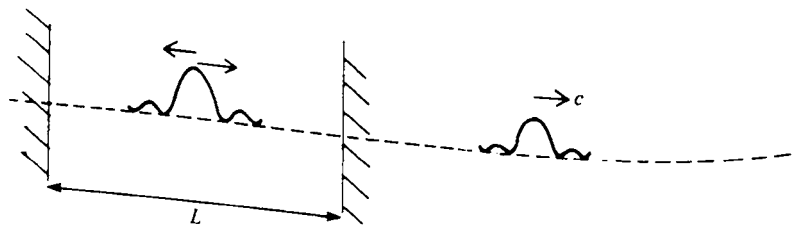


FIG. 6.8 Packet of energy resulting from the mode locking of N modes bouncing to and fro between the laser mirrors. Each time the packet is incident on the output mirror, a 'mode-locked' pulse is emitted.

very briefly every $2L/c$ seconds (corresponding to the cavity round trip time of the wave packet mentioned above). Now if the wave packet is exactly as long in time as the shutter stays open, and if it arrives exactly when the shutter is open, it will be unaffected by the presence of the shutter. Any parts of the wave packet, however, that arrive before the shutter opens or after it closes will be eliminated. Thus, the phase relationships of the oscillating modes are continuously restored by the periodic operation of the shutter.

The electro-optic or acousto-optic modulators discussed in Chapter 3 can be used as shutters, giving rise to mode-locked pulses from an Nd:YAG laser, for example, of about 50 ps duration. In Nd:glass lasers, on the other hand, which generate a very large number of modes because of the broad laser transition line, the pulses can be less than 1 ps duration (see Problem 6.7).

6.3.2 Passive mode locking

Mode locking can also be accomplished by using certain dyes whose absorption decreases with increasing irradiance as shown in Fig. 6.9. Materials exhibiting this behaviour are called *saturable absorbers*. A dye is chosen which has an absorption band at the lasing transition frequency. Initially, at low light levels, the dye is opaque owing to the large number of unexcited molecules which can absorb the light. As the irradiance increases, however, more and more of the excited states are populated until eventually all of them are filled so that the dye becomes transparent. The dye is now said to be *bleached*.

The growth of the mode-locked pulses can be envisaged as follows. Initially, the laser medium emits spontaneous radiation which gives rise to incoherent fluctuations in the energy density within the cavity. Some of these fluctuations, which can be of short duration, may be amplified by the laser medium and grow in irradiance to such an extent that the peak part of the fluctuation is transmitted by the saturable absorber with little attenuation. The low power parts of the fluctuation, however, are much more strongly attenuated and thus a high power pulse can grow within the cavity providing the dye can recover in a time short

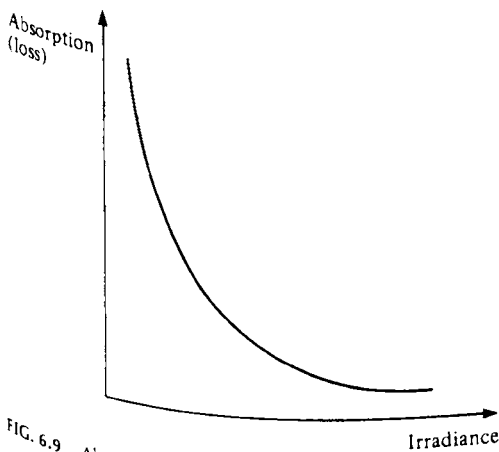


FIG. 6.9 Absorption as a function of incident light irradiance for a saturable absorber.

compared with the duration of the pulse. Because of the non-linear behaviour of the dye, the shortest and most intense fluctuations grow at the expense of the weaker ones. With careful adjustment of the concentration of the dye within the cavity, an initial fluctuation may grow into a narrow pulse 'bouncing' to and fro within the cavity producing a periodic train of mode-locked pulses.

Saturable absorbers provide a simple, inexpensive and rugged method of mode locking high power lasers such as Nd:glass and ruby; the so-called 9740 or 9860 dye solutions and cryptocyanine may be used as the saturable absorber for Nd:glass and ruby respectively. When a saturable absorber is used to mode-lock a laser, the laser is simultaneously Q -switched (section 6.4). The result is the production of a series of narrow (≈ 10 ps) mode-locked pulses contained within an envelope which may be several hundred nanoseconds long. The peak power within the individual pulses may be enormous because of their very short duration.

6.4

 Q -switching

Q -switching is another technique for obtaining short, intense bursts of oscillation from lasers. Single high power pulses can be obtained by introducing time- or irradiance-dependent losses into the cavity. The effects of such losses can be interpreted in terms of the 'spiking' oscillations discussed in section 5.10.1. If there is initially a very high loss in the laser cavity, the gain due to population inversion can reach a very high value without laser oscillations occurring. The high loss prevents laser action while energy is being pumped into the excited state of the medium. If, when a large population inversion has been achieved, the cavity loss is suddenly reduced (i.e. the cavity Q is switched to a high value), laser oscillations will suddenly commence. On Q -switching, the threshold gain decreases immediately (to the normal value associated with a cavity of high Q) while the actual gain remains high because of the large population inversion. Owing to the large difference between the actual and the threshold gain, laser oscillations within the cavity build up very rapidly and all of the available energy is emitted in a single large pulse. This quickly depopulates the upper lasing level to such an extent that the gain is reduced below threshold and the lasing action stops. The time variation of some of the laser parameters during Q -switching is shown schematically in Fig. 6.10. Q -switching dramatically increases the peak power obtainable from lasers.

In the ordinary pulsed mode, the output of an insulating crystal laser such as the Nd:YAG consists of many random 'spikes' of about $1\ \mu\text{s}$ duration with a separation of about $1\ \mu\text{s}$; the length of the train of spikes depends principally on the duration of the exciting flash-tube source, which may be about 1 ms. Peak powers within the 'spikes' are typically of the order of kilowatts. When the laser is Q -switched, however, the result is a single 'spike' of great power, typically in the megawatt range, with a duration of 10–100 ns. It should be noted that, although there is a vast increase in the peak power of a Q -switched laser, the total energy emitted is less than in non- Q -switched operation owing to losses associated with the Q -switching mechanism.

Q -switching is carried out by placing a closed shutter (i.e. the Q -switch) within the cavity, thereby effectively isolating the cavity from the laser medium. After the laser has been pumped, the shutter is opened so restoring the Q of the cavity. A little thought reveals that

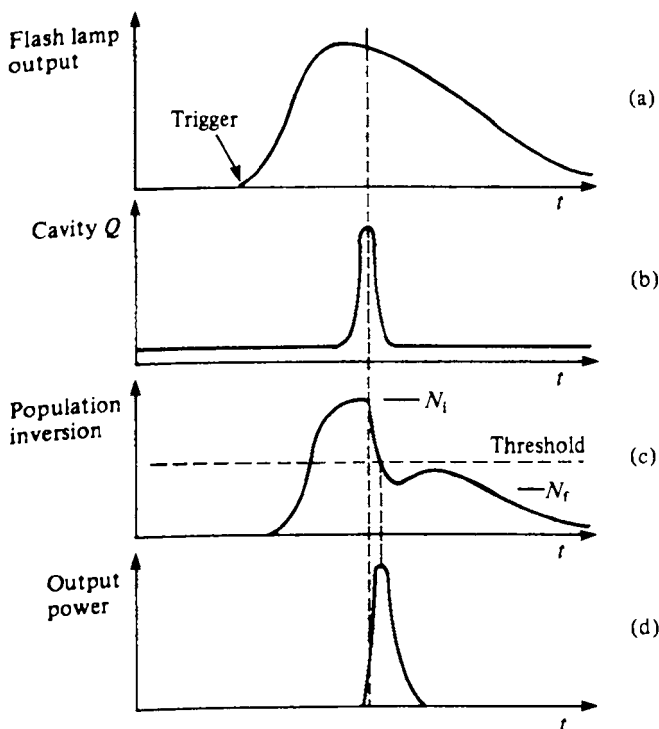


FIG. 6.10 Schematic representation of the variation of the parameters – (a) flash lamp output, (b) cavity Q , (c) population inversion and (d) output power – as a function of time during the formation of a Q -switched laser pulse.

there are two important requirements for effective Q -switching. These are:

1. The rate of pumping must be faster than the spontaneous decay rate of the upper lasing level, otherwise the upper level will empty more quickly than it can be filled so that a sufficiently large population inversion will not be achieved.
2. The Q -switch must switch rapidly in comparison with the build-up of laser oscillations, otherwise the latter will build up gradually and a longer pulse will be obtained so reducing the peak power. In practice, the Q -switch should operate in a time less than 1 ns.

6.4.1 Methods of Q -switching

6.4.1.1 Rotating mirror method

This method, which was the first to be developed, involves rotating one of the mirrors at a very high angular velocity so that the optical losses are large except for the brief interval in each rotation cycle when the mirrors are very nearly parallel. Just before this point is reached a trigger mechanism initiates the flash lamp discharge to pump the laser (here assumed to

be of the insulating crystal type). As the mirrors are not yet parallel, population inversion can build up without laser action starting. When the mirrors become parallel, Q -switching occurs allowing the Q -switched pulse to develop as illustrated in Fig. 6.10.

It should be pointed out that the repetition rate of laser firing is determined by control of the flash lamp and not by the speed of rotation of the mirror, which may be as high as 60 000 rev min⁻¹. If the laser fired every revolution, the repetition rate would be about 1000 times per second, a rate which is prohibitive in insulating crystals owing to the excessive heating of the laser rod which would occur. Although rotating-mirror-type Q -switches are cheap, reliable and rugged, the method suffers from the major disadvantage of being slow. This results in an inefficient production of Q -switched pulses with lower peak power than can be produced by other methods.

EXAMPLE 6.2 Energy of Q -switched pulses

We may estimate the energy of the output pulses from a Q -switched laser as follows. We assume that the population inversion is N_i before the cavity is switched and that it falls to N_f at the end of the pulse (as shown in Fig. 6.10c).

The total energy emitted in the pulse is thus $E = \frac{1}{2}h\nu_{21}(N_i - N_f)V$, where V is the volume of the laser medium. The factor $\frac{1}{2}$ appears because population inversion changes by two units each time a photon is emitted (i.e. the population of the upper level decreases by one while that of the lower level increases by one).

In a typical laser, N_i may be approximately 10^{24} m^{-3} , and assuming $N_f \ll N_i$ and that the laser frequency is $5 \times 10^{14} \text{ Hz}$ and its volume is 10^{-5} m^3 we have that the energy of the pulses is

$$E = \frac{1}{2}(6.63 \times 10^{-34}) \times (5 \times 10^{14}) \times 10^{24} \times 10^{-5} \approx 1.7 \text{ J}$$

It can be shown (ref. 6.6) that the peak power, that is the greatest rate of change of population inversion and hence of photon emission, occurs when the population inversion drops to the threshold inversion N_{th} . To estimate the average power in the pulse, we need to evaluate the pulse duration. We imagine that the Q -switched pulse oscillates to and fro between the laser mirrors and that each time it strikes one of the mirrors a fraction $(1 - R)$ of its energy is lost by transmission. The pulse will then make $1/(1 - R)$ passes along the length of the cavity, which it accomplishes in a time $[1/(1 - R)](L/c)$. This is often referred to as the *cavity lifetime* t_c and may be taken as the duration of the pulse. The power of the pulse is then approximately $P = E/t_c$, which, as $E = \frac{1}{2}h\nu_{21}(N_i - N_f)V$ (see Example 6.2), we can write as

$$P = \frac{(N_i - N_f)h\nu_{21}Vc(1 - R)}{2L} \quad (6.6)$$

EXAMPLE 6.3 Power in Q -switched pulses

Using the data given in Example 6.2 we can estimate the power in the Q -switched pulses from a laser with a cavity length of 0.1 m and a mirror reflectance of 0.8.

The cavity lifetime $t_c = L/(1-R)c = 1.7$ ns. Then the pulse power is given by $E/t_c = 1.7/1.7 \times 10^{-9} = 10^9$ W.

In practice, owing to losses associated with the Q -switch, the actual power in the pulse would probably be nearer 10^8 W.

6.4.1.2 Electro-optic Q-switching

The electro-optic, magneto-optic and acousto-optic modulators described in Chapter 3 can be used as fast Q -switches. If a Pockels cell, for example, is used and the laser output is not naturally polarized, then a polarizer must be placed in the cavity along with the electro-optic cell as shown in Fig. 6.11.

A voltage is applied to the cell to produce a quarter-wave plate which converts the linearly polarized light incident on it into circularly polarized light. The laser mirror reflects this light and in so doing reverses its direction of rotation so that on repassing through the electro-optic cell it emerges as plane polarized light, but at 90° to its original direction of polarization. This light is therefore not transmitted by the polarizer and the cavity is 'switched off'. When the voltage is reduced to zero, there is no rotation of the plane of polarization and Q -switching occurs. The change of voltage, which is synchronized with the pumping mechanism, can be accomplished in less than 10 ns and very effective Q -switching occurs.

Alternative arrangements using Kerr cells and acousto-optic modulators are available. In the case of the acousto-optic modulator, application of an acoustical signal to the modulator deflects some of the beam out of the cavity (see Fig. 3.21), thereby creating a high loss.

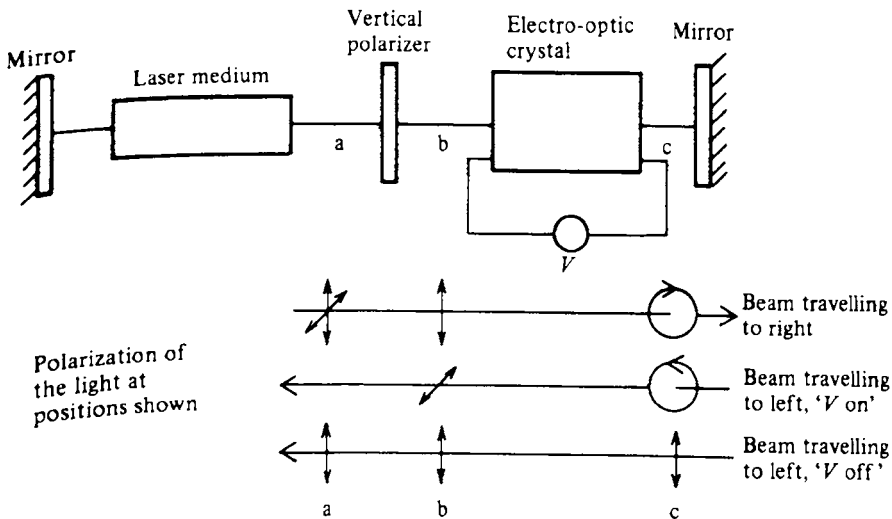


FIG. 6.11 Electro-optic crystal used as a Q -switch. With the voltage V on, the electro-optic crystal acts as a quarter-wave plate and converts the vertically polarized light at b into circularly polarized light at c . The reflected light is converted to horizontally polarized light and eliminated by the polarizer so that the cavity Q is low. With V off, the crystal is ineffective and the cavity Q is high.

When the sound wave is shut off, Q -switching occurs as before. Acousto-optic devices are often used when the laser medium is continuously pumped and repetitively Q -switched, as is frequently the case with Nd:YAG and CO₂ lasers.

6.4.1.3 Passive Q -switching

Passive Q -switching may be accomplished by placing a saturable absorber (bleachable dye) of the type mentioned in section 6.3.2 in the cavity. At the beginning of the excitation flash the dye is opaque, thereby preventing laser action and allowing a larger population inversion to be achieved than would otherwise be the case. As the light irradiance within the cavity increases, the dye can no longer absorb, that is it bleaches and Q -switching occurs. Passive Q -switching has the great advantage of being extremely simple to implement involving nothing more than a dye in a suitable solvent held in a transparent cell. Suitable dyes include cryptocyanine for ruby lasers and sulfur hexafluoride for CO₂ lasers.

As we mentioned in section 6.3.2, lasers that use a saturable absorber for Q -switching are also mode locked if the dye, once bleached, recovers in a time short compared with the duration of the mode-locked pulses.

6.5

Laser applications

In the time which has elapsed since Maiman first demonstrated laser action in ruby in 1960, the applications of lasers have multiplied to such an extent that almost all aspects of our daily lives are touched upon, albeit indirectly, by lasers. They are used in many types of industrial processing, engineering, metrology, scientific research, communications, holography, medicine, and for military purposes. It is clearly impossible to give an exhaustive survey of all of these applications and the reader is referred to the selection of texts and journals given in ref. 6.7. Rather than attempt the impossible, we discuss the properties of laser radiation which make it so useful and reinforce this discussion by brief mention of some appropriate applications. In addition, a rather more detailed description of one or two selected applications is given in sections 6.6 and 6.7.

6.5.1 Properties of laser light

In considering the various properties of laser light we must always remember that not all of the different types of laser exhibit these properties to the same degree. This may often limit the choice of laser for a given application.

6.5.1.1 Directionality

Perhaps the most arresting property of laser light is its directionality. Apart from semiconductor junction lasers, lasers emit radiation in a highly directional, collimated beam with a low angle of divergence. This is important because it means that the energy carried by the laser beam can be collected easily and focused onto a small area. For conventional sources,

where the radiation spreads out into a solid angle of 4π sr, efficient collection is almost impossible, while for lasers the beam divergence angle is so small that efficient collection is possible even at large distances from the laser.

The extent of beam divergence is set by diffraction (section 1.2.4). This is a fundamental physical phenomenon, rather than an engineering limit that can be improved by better optical design. The angle of divergence in radians at the diffraction limit is given by θ , where

$$\theta = \mathcal{H}\lambda/D \quad (6.7)$$

D is the diameter of the aperture through which the beam emerges, and \mathcal{H} is a numerical factor of the order of unity. The precise value of \mathcal{H} depends on the nature of the beam. For example, a TEM_{00} beam has a Gaussian profile (section 5.9.2), and Fig. 6.12 illustrates how the beam diverges outside the laser cavity. We see that the beam divergence angle θ tends asymptotically to the value $\sin^{-1}(w/z)$. Since from eq. (5.37) we also have that $w = z\lambda/\pi w_0$, it follows that at relatively large distances from the laser cavity $\theta = \sin^{-1}(\lambda/\pi w_0)$. Now $\lambda/\pi w_0$ is usually much less than unity, so that we finally obtain $\theta \approx \lambda/\pi w_0$. Assuming that we may associate the minimum beam diameter $2w_0$ with an aperture diameter D , we see that this result agrees with the general divergence equation (6.7) with the parameter \mathcal{H} taking the value $2/\pi$. The beam divergence tends to increase with increasing power output and mode content. Table 6.1 gives some typical beam divergence angles.

The beam may be further collimated by passing it in the reverse direction through a telescope as illustrated in Fig. 6.13. The beam is enlarged by the factor f_2/f_1 and hence the divergence, which is inversely proportional to the beam diameter, is decreased by the factor f_1/f_2 . The ratio of the beam diameters before and after the collimator is given by

$$\frac{D_1}{D_2} = \frac{f_1}{f_2} = \frac{\theta_2}{\theta_1}$$

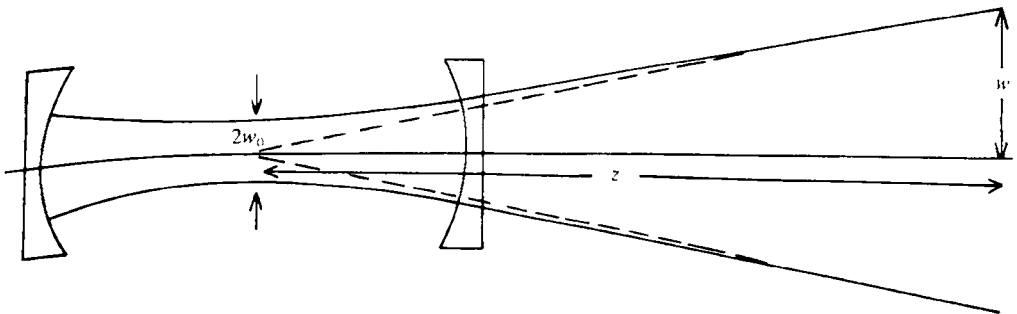


FIG. 6.12 Divergence of a laser beam outside the laser cavity.

TABLE 6.1 Typical laser beam divergence angles

Laser	He-Ne	Ar	CO ₂	Ruby	Nd:glass	Dye	GaAs
Beam divergence (mrad)	0.5	0.8	2	5	5	2	20 × 200

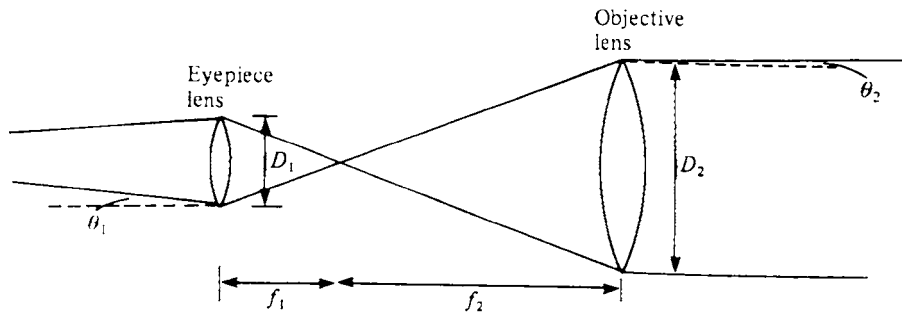


FIG. 6.13 Schematic diagram showing the collimation of a laser beam using a telescope having an eyepiece lens of diameter D_1 and focal length f_1 and an objective lens of diameter D_2 and focal length f_2 . The beam width is enlarged by the factor $D_2/D_1 = f_2/f_1$ and the divergence angle is decreased by the factor f_1/f_2 .

The directional nature, in particular, of gas laser beams readily lends them to applications involving accurate alignment, including civil engineering projects such as drainage and tunnel boring, surveying and the assembly of large aircraft and ships, etc. (for additional references see ref. 6.7e, pp. 245–6).

EXAMPLE 6.4 Beam collimation

We may calculate the reduction in the divergence of a laser beam which is collimated by passing it through a telescope with an objective lens to eyepiece lens focal length ratio of 30:1. Let us consider an He–Ne laser with a plasma tube diameter of 3 mm.

The divergence θ of the beam from the laser is given approximately by λ/D . Therefore $\theta_1 \approx 633 \times 10^{-9} / (3 \times 10^{-3}) \approx 2.1 \times 10^{-4}$ rad (or 0.7 minutes of arc).

Hence, after collimation the angle of divergence will be reduced by a factor of 30, to $\theta_2 \approx 7 \times 10^{-6}$ rad (or 1.4 seconds of arc).

6.5.1.2 Linewidth

Laser light is potentially extremely monochromatic but, as we saw in Chapter 5, the spectral content of the laser radiation may extend over almost as wide a range as the fluorescent linewidth of the laser medium. In other words, although the linewidth of an individual cavity mode may be extremely small there may be many modes present in the laser output. We saw in section 6.1 how single mode operation and frequency stabilization can be achieved. The high spectral purity of laser radiation leads directly to applications in basic scientific research including photochemistry, luminescence excitation spectroscopy, absorption and Raman spectroscopy and also in communications. Many other applications also depend, in part, on this property (ref. 6.7e).

6.5.1.3 Beam coherence

One of the characteristics of stimulated emission is that the stimulated wave is in phase with

the stimulating wave; that is, the spatial and temporal variation of the electric field of the two waves are the same. Thus in a 'perfect' laser we would expect the electric field to vary with time in an identical fashion for every point on the beam cross-section. Such a beam would have perfect *spatial coherence*. Another related property is *temporal coherence*, which refers to the relative phase relationship of the electric field at the same place as a function of time. If the phase changes uniformly with time, then the beam is said to show perfect temporal coherence. These ideas are illustrated in Fig. 6.14.

Coherence is often specified in terms of the *mutual coherence function* $\gamma_{12}(\tau)$ (see ref. 6.8). This quantity, which is in fact a complex number, is a measure of the correlation between

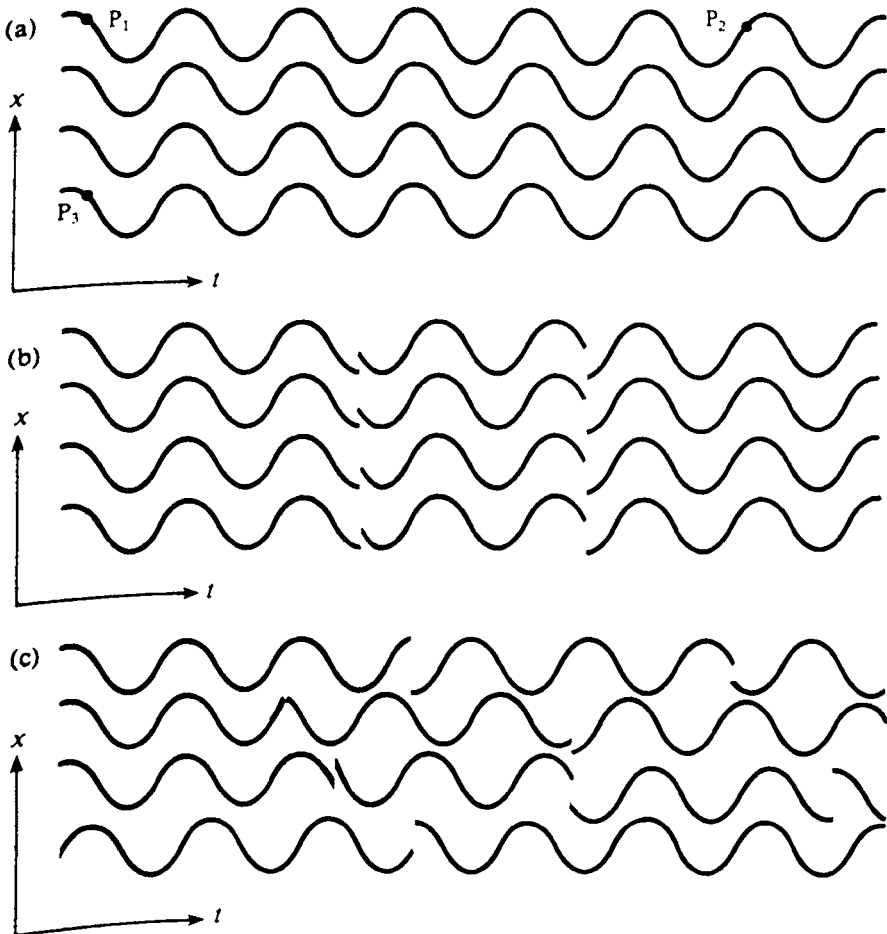


FIG. 6.14 Illustration of coherence. (a) A perfectly coherent beam in which all the constituent waves are in phase at all times. (b) A beam which is spatially coherent but which exhibits only partial temporal coherence. This is because the waves simultaneously change their phase by an identical amount every few oscillations. (c) An almost completely incoherent beam where the phases of each wave change randomly at random times. Note, however, that even in this case some small degree of temporal coherence remains, since over very short intervals the phases are to some extent predictable.

the light wave at two points P_1 and P_2 along the direction of propagation at different times t and $t + \tau$ (temporal coherence), and at points P_1 and P_3 on a plane perpendicular to the direction of propagation at the same time (spatial coherence) (see the representative points on Fig. 6.14a). It has an absolute value between 0 and 1. When it has the value zero the light is completely incoherent, while a value of unity implies complete coherence. Although these extreme values are never achieved in practice, the light from a gas laser operating in a single transverse mode has a value quite close to unity.

Two useful quantities that are related to temporal coherence are the coherence time and the coherence length. To understand these we consider what happens when we take a beam, split it into two equal parts, let the two components travel different distances and then recombine them to form interference fringes as in the Michelson interferometer described in section 6.6.1. Interference effects will only be observed if the path difference of the two beams is such that they are still coherent when they recombine. Light beams from 'real life' sources cannot be represented by infinitely long wave trains, so there is a limit to the path difference. Disregarding lasers for a moment, let us consider a group of atoms undergoing spontaneous emission. Each atom emits radiation independently of the other and does so for only a finite time. This is often because the emission process is perturbed in some way, for example during a collision with a neighbouring atom as described in section 5.7. Thus each atom generates a finite-length wave train and for simplicity we regard the wave trains from all the atoms as having the same length (L_c). Since, however, the atoms are emitting spontaneously, the wave trains are not in phase with each other. If now we pass such a beam into the Michelson interferometer, each individual wave train will be split into two, and the two split wave trains will be able to interfere *provided* the path difference does not exceed L_c . We refer to L_c as the *coherence length*. If the path difference does exceed L_c , then the two halves of each wave train cannot overlap in time when they are recombined, and hence they cannot interfere. In addition, when the path difference lies between zero and L_c , then only a part of each wave train can take part in interference (Fig. 6.15). This implies that as we increase

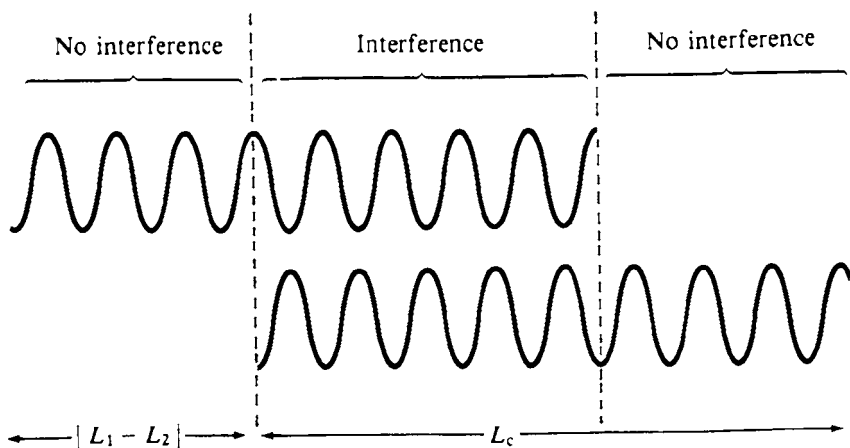


FIG. 6.15 When two identical wave trains of length L_c which have travelled different distances (L_1 and L_2) are recombined, they can only interfere over a length $L_c - |L_1 - L_2|$.

the path difference from zero up to L_c , the irradiance fluctuations corresponding to constructive and destructive interferences will gradually diminish and die out altogether at L_c .

In reality of course, the wave trains are not all the same length but are distributed in some way about a mean value. The conclusion that the irradiance fluctuations will gradually reduce, however, still remains valid. Similarly the exact definition of coherence length is somewhat more involved, but we may still take it to be the path difference at which the irradiance fluctuations die out.

The *coherence time* t_c may then be defined as the time taken for the source to emit a wave train of length L_c . Thus

$$t_c = L_c/c \quad (6.8)$$

where c is the velocity of light. In fact it can be shown (ref. 6.8) that the coherence time is related to the linewidth of the emission ($\Delta\nu$) via the equation

$$t_c = \frac{1}{\Delta\nu} \quad (6.9)$$

Although we started this discussion of coherence by considering spontaneous emission, the ideas of coherence length and coherence time apply equally well to laser radiation.

EXAMPLE 6.5 Coherence lengths of conventional and laser radiation sources

We consider first of all the light emitted from a low pressure sodium lamp. A typical linewidth of the sodium D lines (both lines taken together) at $\lambda_0 = 589 \text{ nm}$ is $5.1 \times 10^{11} \text{ Hz}$. Equation (6.9) then gives $t_c = 2 \times 10^{-12} \text{ s}$ and from eq. (6.8), $L_c = 0.6 \text{ mm}$. We may contrast these values with those applicable to an He-Ne laser. If many modes are operating, then the linewidth is about 1500 MHz giving a coherence length of 0.2 m. However, if the laser is operating in a single mode stabilized to 1 MHz, then the coherence length is some 1500 times greater, that is 300 m.

Thus for there to be coherence at the points P_1 and P_2 in the direction of propagation, the separation of P_1 and P_2 should be less than L_c .

Consider now a perfectly collimated beam of monochromatic light; then the phase is the same at any point on a surface perpendicular to the beam, and the points P_1 and P_3 may be separated by quite large distances. On the other hand, for a non-monochromatic, incompletely collimated beam the points P_1 and P_3 can only be a fraction of a millimetre apart if there is to be any coherence between them. We sometimes define a transverse coherence length, L_t , which represents the distance perpendicular to the main propagation direction over which the phases at two points remain correlated. If there is coherence between the points P_1 and P_3 then interference effects should result if the light from these points can be brought together. In principle, this can be accomplished using the Young's slits apparatus described in section 1.2.3.

Example 6.5 illustrates how the presence of several modes can dramatically reduce the temporal coherence of a laser, and the same is also true of the spatial coherence. A well-stabilized laser emitting a single transverse mode exhibits almost perfect spatial coherence

across the entire beam cross-section. It has been possible to expand the TEM₀₀ mode of gas lasers to give transverse coherence lengths of several metres for holographic purposes (see section 6.7). On the other hand, if a number of transverse modes are present the spatial coherence is considerably reduced.

While the coherence of the output from CW gas lasers can be very high, that from pulsed lasers is usually somewhat smaller. This is because the temporal coherence may be limited by the duration of the spikes within the laser pulse or by shifts in frequency during emission. Thus, ruby lasers emit spikes that are typically 0.6 μs in duration and exhibit coherence times of about 0.1 μs. The coherence lengths of some common lasers are given in Table 6.2.

Coherence is important in any application where the laser beam will be split into parts. These include interferometric measurement of distance (section 6.6) and deformation, where the light is split into parts that traverse different distances, and holography, where the light beams traverse different paths which may be approximately equal but which may have spatially different distributions.

One of the most striking characteristics of laser light reflected from rough surfaces is its speckled or granular appearance. This is the result of a random interference pattern formed from the contributions of light reflected from neighbouring portions of the surface. In some regions, these contributions will interfere constructively, while in others they interfere destructively. This behaviour is a direct consequence of the high coherence of laser light.

In many applications, for example holography, the speckle pattern is a nuisance, though the phenomenon is finding application in a number of areas including metrology and vibration analysis (ref. 6.9).

6.5.1.4 Brightness

The primary characteristic of laser radiation is that lasers have a higher brightness than any other light source. We define *brightness* as the power emitted per unit area per unit solid angle (sometimes the term *specific brightness*, i.e. brightness per unit wavelength range, $W\ m^{-2}\ sr^{-1}\ \Delta\lambda^{-1}$, is used). (In radiometry this unit – brightness – is called the radiance, see section 1.5, but in laser work the term brightness as defined here is used.) The relevant solid angle is that defined by the cone into which the beam spreads. Hence, as lasers can produce high levels of power in well-collimated beams, they represent sources of great brightness.

TABLE 6.2 Summary of coherence lengths of some common lasers

Laser	Typical coherence length
He-Ne single transverse, single longitudinal mode	Up to 1000 m
He-Ne multimode	0.1 to 0.2 m
Argon multimode	0.02 m
Nd:YAG	10 ⁻² m
Nd:glass	2 × 10 ⁻⁴ m
GaAs	1 × 10 ⁻³ m
Ruby:	
for whole output pulse	10 ⁻² m
within a spike-forming part of the pulse	≪ c times spike length, i.e. ≪30 m

The brightness is also affected by the presence of additional modes, for often as the laser power is increased the number of modes increases but the brightness remains almost constant. Typical values of brightness are: for an He-Ne laser, $10^{10} \text{ W m}^{-2} \text{ sr}^{-1}$; for a Q-switched ruby laser, $10^{16} \text{ W m}^{-2} \text{ sr}^{-1}$; and for an Nd:glass laser followed by amplifiers, $10^{21} \text{ W m}^{-2} \text{ sr}^{-1}$ has been achieved. For comparison, the brightness of the sun is about $1.3 \times 10^6 \text{ W m}^{-2} \text{ sr}^{-1}$!

High brightness is essential for the delivery of high power per unit area to a target; this in turn depends on the size of the spot to which the beam can be focused.

6.5.1.5 Focusing properties of laser radiation

The minimum spot size to which a laser beam can be focused is determined by diffraction. We have seen in section 5.9 that all laser beams possess a degree of divergence, even though in some cases this might be quite small. We can 'reverse' this divergence by inserting a suitable focusing lens into the beam; the lens brings the laser beam to a focus at a distance nearly equal to f , its focal length (section 1.3.2).

If the laser beam has a radius w_L at the lens then by analogy with eq. (5.7), we may write

$$w_L = \frac{\lambda f}{\pi r_s}$$

where r_s represents the focused beam radius, or 'spot size'. Hence

$$r_s = \frac{\lambda f}{\pi w_L} \quad (6.10)$$

If the laser beam exactly fills the lens aperture we can set $2w_L = D$, where D is the diameter of the lens, so that

$$r_s = \frac{2\lambda f}{\pi D} = \frac{2}{\pi} \lambda F \quad (6.10a)$$

where $F (=f/D)$ is the *F number* of the lens. It is impracticable to work with F numbers much smaller than unity so that r_s is of the order of λ . It is interesting to note from eq. (6.10) that the larger the beam diameter at the focusing lens the smaller is the focused spot radius r_s . For this reason the laser beam is often passed through a beam expander before being focused onto the workpiece.

Once again the presence of a complicated mode structure in the beam is deleterious in that, in this case, the focused spot size is much larger and the power density (irradiance) is correspondingly much smaller for a given laser power. Again, if the beam divergence is large the power density is reduced. Nevertheless, doped insulator lasers, because of the very high peak powers they generate, can easily produce very high irradiances. A focal area of 10^{-7} m^2 is typical for such lasers, giving rise to typical average irradiances of 10^9 W m^{-2} and peak irradiances of 10^{12} W m^{-2} .

EXAMPLE 6.6 Focused power densities of laser radiation

We consider a 10 mW He–Ne laser focused by a lens with an F number of 1.

From eq. (6.10a) the radius of the focused spot is

$$r_s = 632.8 \times 10^{-9} \times (2/\pi) \text{ m}$$

Hence the power per unit area is equal to

$$\frac{10 \times 10^{-3} \times (\pi)}{(632.8 \times 10^{-9} \times 2)^2} = 2 \times 10^{10} \text{ W m}^{-2}$$

As we shall see in section 6.8 such high irradiances lead to the use of lasers in the drilling, cutting, welding and heat treatment of a large number of different materials (ref. 6.10). In certain applications, for example the micromachining of electronic components, good focusing is required and hence we would wish to use a short focal length lens. This, however, may be impracticable on a production line because of the limited depth of focus (or field). We must provide sufficient *depth of focus* to allow for vibrations and inaccuracy in positioning in the vertical sense. The depth of focus is the distance that we can move the workpiece away from the position of minimum beam radius and still have an acceptably small spot of light. From eq. (5.34) we know that the beam radius varies with distance in a parabolic manner. Rearranging this equation we have

$$z = \frac{\pi w_0^2}{\lambda} \left[\left(\frac{w(z)}{w_0} \right)^2 - 1 \right]^{1/2} \quad (6.11)$$

Thus if we know the maximum variation in spot size that is acceptable, that is the maximum value of $w(z)/w_0$, then eq. (6.11) enables the maximum tolerable variation in z to be calculated. We note that the smaller we make the minimum beam radius, the smaller is the depth of focus, and we have to reach an acceptable compromise between sufficiently large depth of focus and small focal area.

EXAMPLE 6.7 Depth of focus for a CO₂ laser beam

We consider a CO₂ laser beam ($\lambda_0 = 10.6 \mu\text{m}$), which is passed through a beam expander to fill the 50 mm diameter aperture of a focusing lens which has a 200 mm focal length. Using eq. (6.10), we can calculate the radius of the focused spot, which we here equate to w_0 in eq. (6.11), that is

$$r_s \equiv w_0 = \frac{10.6 \times 10^{-6} \times 200 \times 10^{-3}}{\pi \times 25 \times 10^{-3}} = 30 \mu\text{m}$$

We suppose that a 10% variation in the spot size can be tolerated, that is $w(L)/w_0 = 1.1$. Hence the depth of focus is given by

$$\frac{\pi(30 \times 10^{-6})^2}{10.6 \times 10^{-6}} [(1.1)^2 - 1]^{1/2}$$

or 0.12 mm.

The selection of a laser for a given application involving laser 'heating' depends very much on the nature of the application. For many cutting tasks it may be advantageous to use a CW laser; for continuous output, the highest powers are produced by CO₂ lasers for which values up to 100 kW have been quoted. For welding operations, a pulsed laser may be preferred. In this case, because of the very short pulses which can be produced, we find that *Q*-switched Nd:glass lasers generating pulses with peak powers of about 10¹¹ W are commercially available.

The focusing properties of laser radiation are also important in low power applications, two of which represent the first laser-based devices to be used by the public at large. The first of these is the 'point of sale' device used to price items in supermarkets and to provide for automatic stock information upgrading. Products have a coded label, consisting of a series of parallel bars of varying widths, placed on them. This is scanned by a laser. The light reflected from the bars is detected, thereby identifying and pricing the product and adding the price to the bill.

The second application is in the preparation and readout of video-disk systems. The information is imprinted on the video-disk in digital form by forming small pits in the surface of the disk with a laser. These pits are subsequently read by a low power laser to provide a video signal for playback on a television set. The system has a number of advantages over tape or stylus pick-up systems including the absence of wear, the high information density that can be accommodated by the closely focused laser beam and the fact that warped disks can be played equally well. Also high quality 'frozen' images can be selected at will and held indefinitely. Similar arrangements are used in compact disc (CD) audio systems, and are becoming increasingly popular for archival and computer storage (ref. 6.11).

6.5.1.6 Tunability

We saw in Chapter 5 and the earlier part of this chapter that some lasers can be tuned to emit radiation over a range of wavelengths. With dye lasers, for example, the range of tunability can be large. Indeed dye lasers can be turned over most of the visible spectrum, and by harmonic generation this range can be extended into the ultraviolet. On the other hand, optical parametric amplifiers used with a primary laser source can upconvert into the range 1–25 μm in the infrared (section 3.9.1).

Laser tunability leads to applications in photochemistry, high resolution and Raman spectroscopy and isotope separation (ref. 6.12).

6.6 Measurement of distance

The main methods of measuring distance using lasers are: (a) interferometry; (b) beam modulation telemetry, and (c) pulse time of flight (ref. 6.13a).

6.6.1 Interferometric methods

We saw in Chapter 1 that if the wavefront from a light source is divided into two parts which then traverse different distances before being recombined, then an interference fringe pattern

is produced. The irradiance distribution in the pattern is characteristic of the point-for-point path differences between the two parts of the beam. Thus, if one of the path lengths is changed the fringe pattern will move across the field of view and the change in path length can be measured in terms of the fringe shift.

The classical method for measuring distance (or changes of distance) in this way is the Michelson interferometer and nearly all other methods are variations of this instrument. The Michelson interferometer, which is shown in Fig. 6.16, consists of a beam splitter, two plane mirrors and an observing telescope. The wavefront from the laser source is divided by the beam splitter B; the two parts then proceed to the plane mirrors M_1 and M_2 and are reflected back to B. Some of the light is reflected by the beam splitter and some is transmitted as shown so that the beam splitter serves to recombine the beams and interference fringes can be seen through the telescope. We may regard the fringes as being produced in the thin film formed between the mirror M_1 and M'_2 which is the reflection of mirror M_2 in B. Thus if M_1 and M'_2 are exactly parallel, that is M_1 and M_2 are exactly perpendicular to each other, a system of circular fringes will be seen as explained in Chapter 1. On the other hand, if one of the mirrors is tilted slightly, then a system of straight line fringes is formed.

We saw in Chapter 1 (eq. 1.23) that for thin film interference a bright fringe is formed when

$$p\lambda_0 = 2D \cos \theta \approx 2D \quad (\text{if } \theta \text{ is small})$$

where D is the optical thickness of the film. Hence if one of the mirrors, M_2 say, is moved D will change and the fringe pattern will move. Specifically, if D changes by $\lambda_0/2$ a complete fringe will pass a reference point in the field of view. Therefore we can measure the

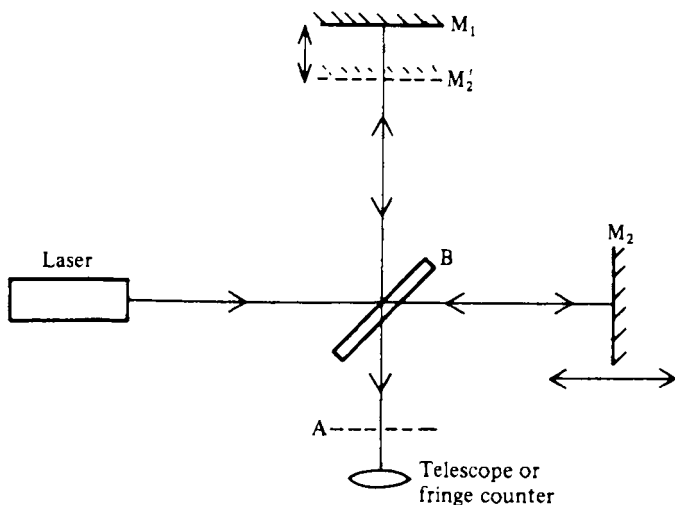


FIG. 6.16 Michelson interferometer: M_1 is a fixed mirror, M_2 is movable and M'_2 is its image in the beam splitter B. Interference fringes can be seen through the telescope, or alternatively the number of fringes crossing an aperture placed in position A can be detected using a photodetector and counted electronically.

distance moved by M_2 in terms of fringe shifts, that is in terms of the wavelength λ_0 of the light used. To measure an unknown distance one simply aligns M_2 with one end and counts the fringe shift as it is moved until it coincides with the other end of the distance being measured.

Optical interferometry predates the laser by many years, of course, but the technique was always limited by the coherence limitations of the light sources available. Distances of a few centimetres could be measured at best. With an He-Ne laser, however, coherence lengths of many metres are available so that, in principle, we can measure up to such distances with an accuracy of a fraction of a wavelength. Fringe displacements of 0.01 of a fringe, equivalent to $\lambda/200$, can be detected. We must remember, however, that the distances measured are optical path lengths which include the refractive index of the air. Changes in refractive index, due to pressure and temperature variations and atmospheric turbulence, result in random fringe shifts and thereby limit the distance which can be measured and the accuracy attainable. Accuracies of about 1 part in 10^6 can be achieved quite readily. In practice the plane mirrors are replaced by cube corner retroreflectors (Fig. 6.17). These have the property of reflecting an incident beam back along a direction parallel to its incident path thereby simplifying the alignment of the instrument. The lateral displacement involved prevents the returning light from entering the laser cavity and thereby creating an undesirable modulation of the laser output. The large number of fringes which cross the field of view is counted electronically using, for example, a silicon photodiode as a detector.

The technique is widely used in machine tool control, length standard calibration and for seismic and geodetic purposes.

6.6.2 Beam modulation telemetry

As we mentioned above, owing to fluctuations in the density of the atmosphere, interferometric distance measuring methods are limited to distances not exceeding about 100 m. For greater distances, techniques involving amplitude modulation of the laser beam are useful. The beam from an He-Ne or GaAs laser is amplitude modulated and projected to

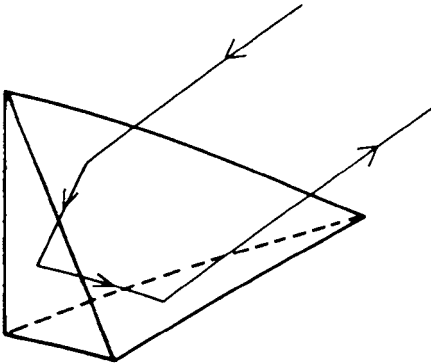


FIG. 6.17 Schematic diagram of a cube corner reflector. The incident light reflects once off each side face and emerges along a path almost exactly parallel to that of the incident light.

the 'target' whose distance is to be measured. The light reflected from the target is received by a telescope and sent to a detector. The phase of the modulation of the reflected beam is different from that of the emitted beam because of the finite time taken for the light to travel to the target and return to the telescope. The phase shift ϕ is given by

$$\phi = \frac{2\pi}{\lambda_0} (2n_g L) \quad (6.12)$$

where L is the target distance and n_g is the group index of refraction of the atmosphere. The value of n_g for the 632.8 nm He-Ne laser wavelength is $n_g = 1.000\,28$ for dry air at 15°C, 760 torr and 0.03% CO₂. Corrections for varying atmospheric temperature and pressure are available. These corrections are difficult to apply for measurements in the field, however, and one must attempt to average n_g over the entire path traversed by the light.

Figure 6.18 shows the schematic diagram of a beam modulation system. The light is amplitude modulated at a given frequency f , collimated and transmitted to the target. Reflected light is collected by the telescope and focused onto the detector (the presence of a retroreflector on the target is a great help). A phase detector compares the relative phase of the reflected beam with that of the original beam.

The phase difference can be written as

$$\phi = (p + q)2\pi$$

where p is an unknown integer and q is a fraction less than unity. The phase comparison gives q but not p . To find p , the measurement must be repeated with different values of the modulation frequency. Having found ϕ then L can be determined from eq. (6.12). The narrow bandwidth of the laser light enables high discrimination against stray light so that the system

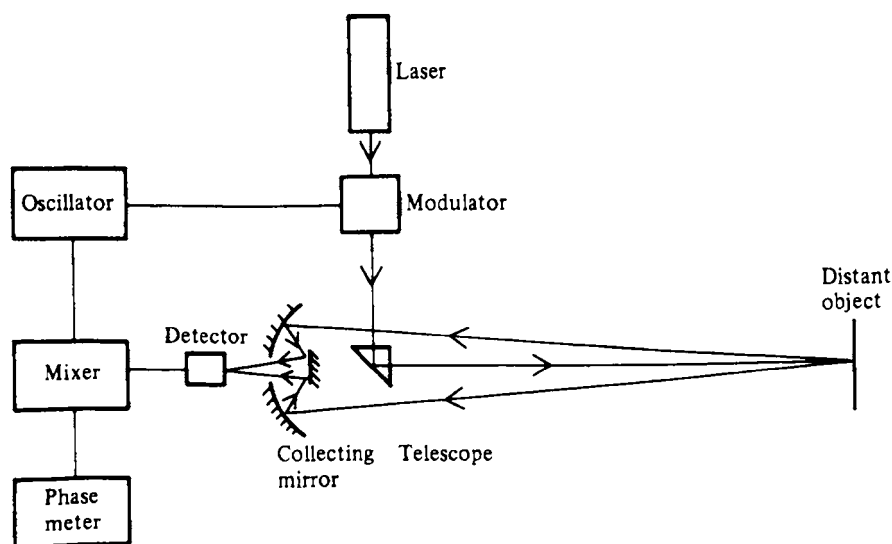


FIG. 6.18 Schematic diagram of a beam modulation distance measurement system.

can be used in daylight with a high signal-to-noise ratio, while the small beam divergence allows a high degree of selectivity in the target being examined.

Beam modulation telemetry units have been developed for which the accuracy is better than 1 mm at distances up to 1000 m and 1 part in 10^6 for greater distances. Such devices have been used for the measurement of large structures such as dams and bridges, and as airborne instruments for land profiling and for geodetic surveying.

6.6.3 Pulse echo techniques

We can measure large distances by timing the round trip transit time for a very short pulse reflected from a distant target. The system consists of a pulsed laser, preferably *Q*-switched, a telescope to collect the reflected light, a photodetector and an accurate timer. The narrowly collimated beam of the laser makes it possible to measure the range to specific targets and the technique has military applications as a range finder. Accuracy of the order of ± 5 m in 10 km has been achieved.

A novel application has been in measuring the distance to the moon. Using retroreflectors left on the surface of the moon during the Apollo 11, 14 and 15 space missions, the lunar distance has been measured to an accuracy of ± 15 cm.

This technique, which is often known as *optical radar* or '*lidar*' (*light detection and ranging*), has been extended to atmospheric studies. By measuring the amount of backscattered light, the presence of air turbulence can be detected and the amounts of various atmospheric pollutants such as CO_2 and SO_2 can be measured (refs 6.12c and 6.13b).

6.7

Holography

Although holography was developed prior to the laser (the first hologram was recorded by Gabor in 1948, ref. 6.14), the requirement of holography for light with a high degree of spatial and temporal coherence has closely linked the development of holography to that of lasers. Holography is a method of recording information from a three-dimensional object in such a way that a three-dimensional image may subsequently be reconstructed; the phenomenon is often known as wavefront reconstruction. A great deal of work has been done on holography and its applications; a selection of texts on the subject is given in ref. 6.15.

Figures 6.19(a) and (b) illustrate the basic principles. A photographic plate is exposed simultaneously to waves of light scattered by the 'object' and to waves of light from a 'reference' source. The reference beam, shown here in Fig. 6.19(a) as a plane-parallel beam, may be of any reproducible form and is derived from the same laser source as the light illuminating the object. Because of their high degree of mutual coherence the two sets of waves produce an interference pattern on the plate, which is recorded in the photographic emulsion and forms a *hologram*.

The photographic plate is now processed and illuminated with only the reference beam present as shown in Fig. 6.19(b). Most of the light from the reference beam passes straight through the hologram; some of it, however, is diffracted by the interference pattern in the emulsion. By the normal diffraction grating equation (eq. 1.28), light of wavelength λ will

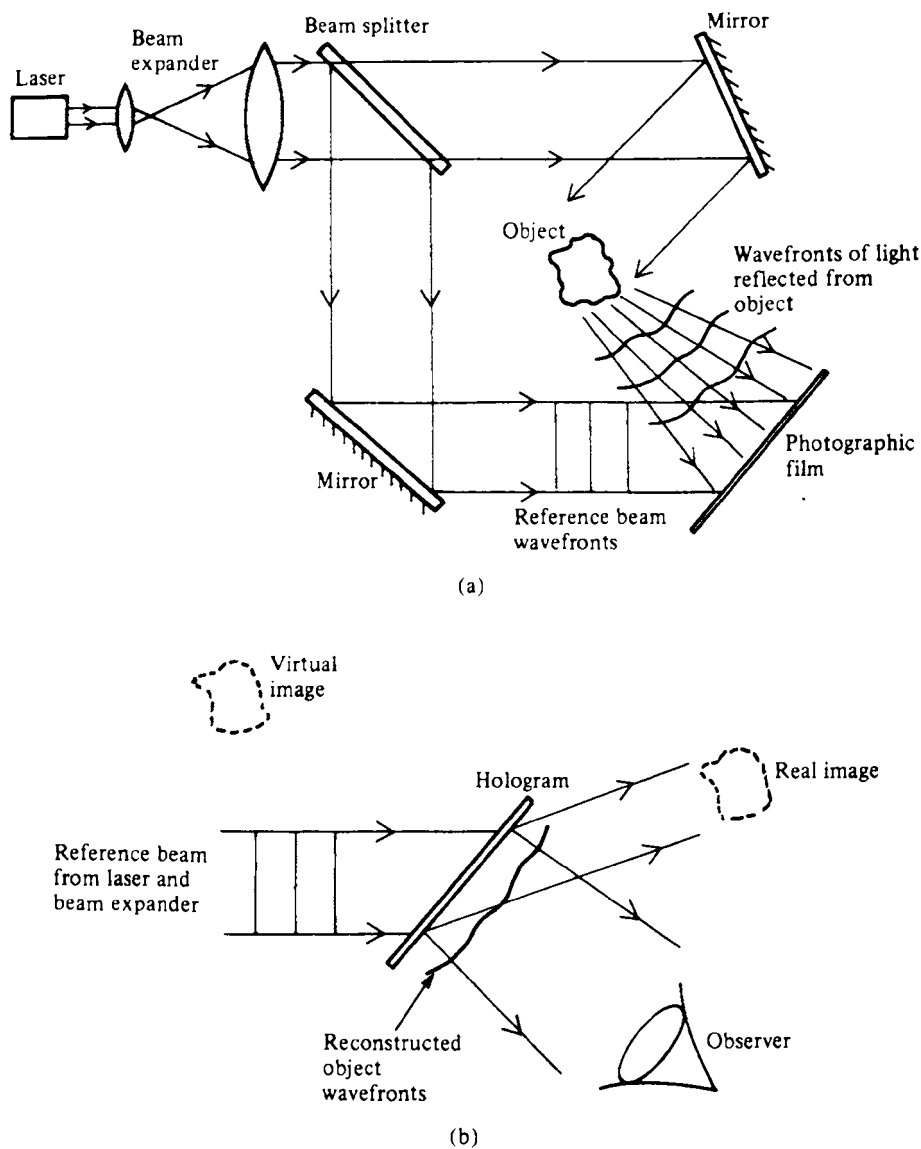


FIG. 6.19 Schematic diagram showing the basic geometry for (a) making a hologram by recording the interference pattern produced by the interference of light reflected from the object, and the reference wavefronts, and (b) reconstruction of the object wavefront. To the observer the reconstructed wavefront appears to be coming from the object itself and a virtual image is seen.

experience constructive interference at angles such that $\lambda = d \sin \theta$, where d is the local fringe spacing of the interference fringes whose exact shape and distribution depends on the shape of the object and the wavefronts reflected from it. Thus the constructive interference of these diffracted waves reconstructs the original wavefronts from the object and to an observer the

wavefronts appear to be coming from the object itself. These wavefronts constitute what is termed the *virtual image*. However, just as a diffraction grating gives diffracted orders on either side of the 'straight-through' position, the hologram generates a second image; this image, which is usually inferior in quality to the virtual image, is called the *real image*.

The hologram serves as a 'window' on the object scene, which has been illuminated by the laser, through which the object can be viewed from different angles. The range of views of the reconstructed object is limited only by the size and position of the hologram and a truly three-dimensional effect is created.

The mathematical analysis of holography is quite complicated, but the following simplified treatment will serve to illustrate the principles involved. We assume that the photographic plate is in the (x, y) plane and that we may represent the electric field of the wavefront reflected from the object in the (x, y) plane at time t by

$$E_{ob} = U_0(x, y)\exp(-i\omega t)$$

where $U_0(x, y)$ is the amplitude, which in general is complex. Similarly, the complex amplitude of the reference beam in the (x, y) plane at the same instant is $U_r(x, y)$. As the object and reference beams are coherent, the irradiance recorded on the photographic plate is given by adding the amplitudes and multiplying by the complex conjugate. Thus the irradiance is

$$I(x, y) = |U_0 + U_r|^2 = (U_0 + U_r)(U_0^* + U_r^*)$$

Therefore

$$I(x, y) = (U_0U_0^* + U_rU_r^*) + (U_0U_r^* + U_0^*U_r) \quad (6.13)$$

The first term in eq. (6.13) is the sum of the individual irradiances; the second term represents the interference which occurs and thus contains information in the form of amplitude and phase modulations of the reference beam.

The plate is now processed to form a transmission hologram. With correct processing, the transmission of the hologram is a constant (say T) times the irradiance function $I(x, y)$ given in eq. (6.13). If the plate is now illuminated by the original reference beam only, the transmitted light will have a complex amplitude U_T which is proportional to U_r times the transmittance of the hologram $TI(x, y)$. Hence we may write

$$\begin{aligned} U_T(x, y) &= U_r TI(x, y) \\ &= T[U_r(U_0U_0^* + U_rU_r^*) + U_r^2U_0^* + U_rU_r^*U_0] \end{aligned} \quad (6.14)$$

As we mentioned above, the hologram behaves like a diffraction grating and produces a direct beam and a first-order diffracted beam on either side of the direct beam. The first term in eq. (6.14) represents the direct beam. The last term is the one of greatest interest; $U_rU_r^*$ is constant so that the last term is essentially U_0 , the object wavefront amplitude. Hence this diffracted beam represents a reconstruction of the wavefront from the original object and it forms the virtual image. The middle term represents the other diffracted beam and forms the real (conjugate) image.

This description can be verified by considering the simple case of an object comprising a single white line on a dark background. The hologram, in this case, turns out to be a simple

periodic (or sine) grating. The zero order of the diffracted light is the direct beam, whereas the first orders on either side comprise the virtual and real images.

6.7.1 Applications of holography

Undoubtedly, the full potential of the holographic technique is still to be realized although a number of applications are now firmly established (ref. 6.16). We shall describe one established and one potential application, namely holographic interferometry and computer memories respectively.

6.7.1.1 Holographic interferometry

The determination of surface contours by conventional interferometry has been restricted to the examination of reflecting surfaces with simple shapes. This restriction is removed by holographic interferometry, which can be used for complicated shapes with diffusely reflecting surfaces. There are a number of recognizably different types of holographic interferometry which we now describe briefly.

Double exposure holographic interferometry is an important industrial process in which very small displacements or distortions of an object can be measured by counting interference fringes. The subject of the investigation is recorded as a hologram and before the plate is processed the subject is moved, distorted by stress or whatever, and a second hologram recorded. After processing, each image can be reconstructed in the usual way. The two sets of wavefronts for the reconstructed images interfere and produce interference fringes over the full range of views of the subject obtainable through the hologram. A typical example of this technique is given in Fig. 6.20 which shows a circular membrane which has been deformed by a uniform pressure. The time between the two records may be anything from a fraction of a microsecond upwards, but the plate and subject must maintain the same relative positions except for the movement under investigation.

A variation of the technique is *real time holographic interferometry*, in which the interference fringes are viewed in real time. A hologram of the subject is recorded as above but in this case the plate is processed and replaced in its original place. The subject is now distorted and interference fringes can be observed through the holographic plate, changing as the distortion of the subject actually occurs. Although real time holographic interferometry provides a sensitive tool for measuring the strains of objects as they actually deform, it suffers from a number of problems. These include the difficulty of replacing the plate *exactly* in its original place and distortion of the photographic emulsion during processing. Figure 6.21 shows the fringe pattern in real time holographic interferometry as the object, a metal bar which is clamped at one end, is stressed.

The third technique, *time-average holographic interferometry*, is particularly useful for examining the spatial characteristics of low amplitude vibrations of an object. In most holographic situations, a general rule is that the object should remain stationary during the period of exposure. In the present case this rule is violated dramatically, for during the exposure the object is moving continuously. The resulting hologram may be regarded as the limiting case of a large number of exposures for many different positions of the surface. The fringes

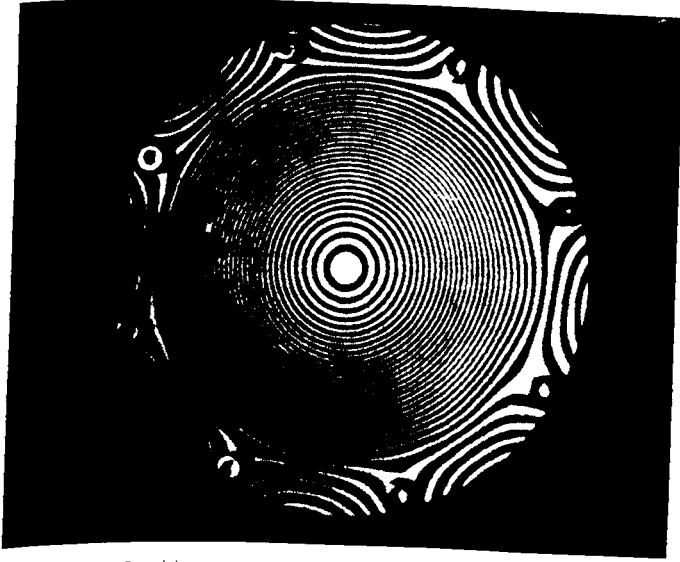


FIG. 6.20 Double exposure holographic interferogram showing the deformation of a circular membrane which has been caused by a uniform pressure. (Photograph courtesy of W. Braga and C. M. Vest, University of Michigan.)

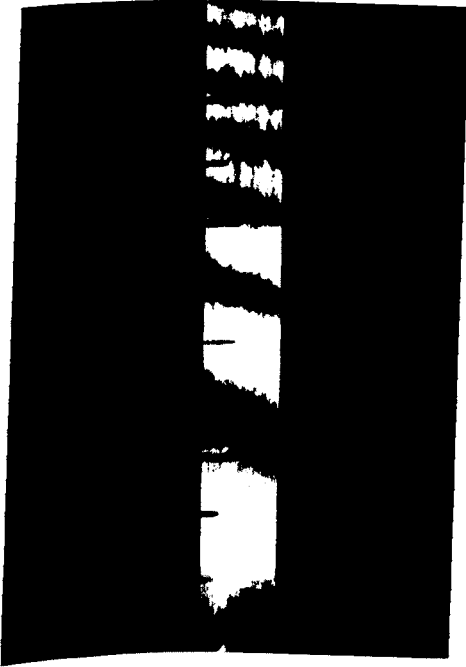


FIG. 6.21 Illustration of real time or single exposure holographic interferometry. Interference of the actual object wave with the reconstructed object wave shows the deformation of the bar. The fact that the fringes are not horizontal indicates that the bar has suffered a twist in addition to bending. (From O'Shea, Callen and Rhodes, *Introduction to Lasers and their Applications* © 1977, Addison-Wesley, Reading, MA. Fig. 7.14. Reprinted with permission.)

produced represent contour lines of equal amplitude of vibration of the surface. The brightest fringes occur at the nodes where the surface remains stationary. Elsewhere, providing the period of exposure covers many vibrations of the surface, there is a variation in irradiance due to the surface motion, with almost zero irradiance at the antinodes. An example of the application of time-average holographic interferometry to the analysis of the vibrations of a turbine blade is shown in Fig. 6.22.

6.7.2 Holographic computer memories

Holographic computer memories are being actively developed because, potentially, they have a very high storage capacity—theoretically up to 10^{10} bits mm^{-3} —with rapid access. Though they are unlikely to replace established technologies such as magnetic disk, tape or optical disk storage, they could be used where very large total storage capacity is required. Holographic storage systems are also interesting in that they involve several of the opto-electronic devices we have described, such as lasers, optical modulators, photodetectors, as well as optical storage media.

A holographic memory records and reads out a large number of bits simultaneously as we can appreciate by considering Fig. 6.23. The information to be stored is formed into a two-dimensional array of bits by a device called a *page composer*. The *page composer* may be thought of as an array of light valves which may be open or closed corresponding to ‘ones’



FIG. 6.22 Holographic reconstructions from time-average holograms showing flexural resonances (A and B) and torsional resonances (C and D) of a turbine blade. (from an article by Robert K. Erf in Robert K. Erf (Ed), *Holographic Non-Destructive Testing*, 1974, Courtesy Academic Press Inc.)

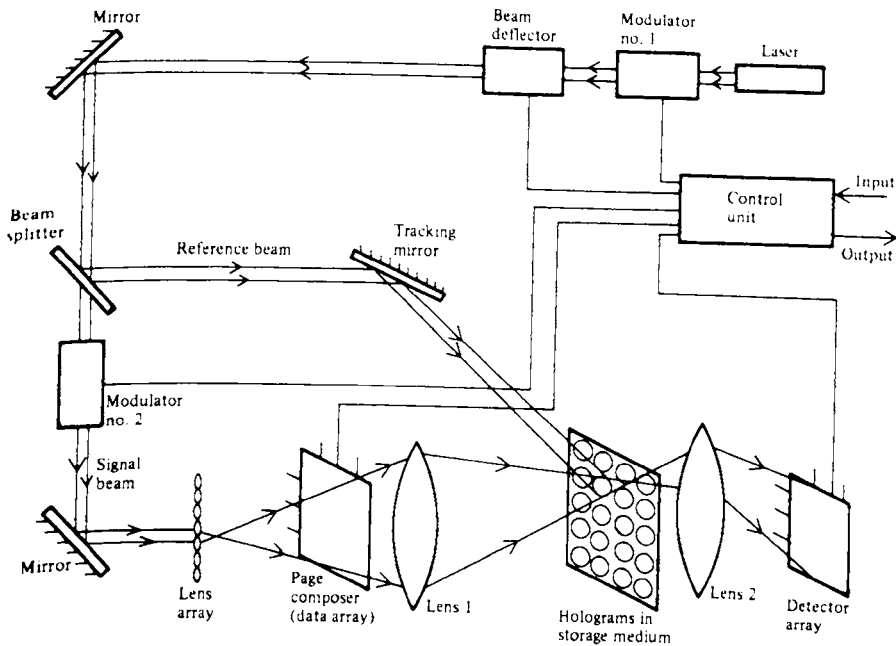


FIG. 6.23 Schematic diagram of a holographic memory system.

and 'zeros' respectively. This array of up to perhaps 10^5 – 10^6 bits is then stored at one time in a particular location of the holographic memory. During recording, the light modulators allow maximum irradiance in both the signal and reference beams. To store a different array as another hologram on the storage medium the beam deflector moves the beams to the appropriate position.

Readout of data occurs when the hologram is addressed only by the reference beam. The first modulator is partly closed to reduce the irradiance of the light reaching the holographic storage medium, while the second modulator is closed to cut off the signal beam. The deflector directs the beam to the hologram to be read via the tracking mirror and an image representing the arrays of 'ones' and 'zeros' is produced. This image is focused by a lens and projected onto the detector array such as a charge-coupled device (CCD) array, which is now available with 2048×2048 detector pixels – see section 7.3.7. Each bit originally stored in the page composer is incident on one photodetector in the array. The stored information is thus converted into electrical signals.

The storage of data in this way offers a number of advantages compared to a bit-orientated memory. The information about a particular array is distributed in a holographic fringe pattern across the entire hologram. Thus the hologram is not sensitive to small scratches or particles of dust which might otherwise cause the loss of bits of data in a bit-orientated memory.

Secondly the information is recovered essentially in parallel. A large number of bits are all read simultaneously allowing a very high readout rate. The requirements on light beam deflection are also reduced. Each position to which the beam is deflected represents about 10^5 bits of data; thus a 10^9 bit memory requires only 10^4 separate locations. This lies within

the capabilities of present light beam deflectors. Addressing can be done using electro-optical deflectors which have access times in the region of microseconds.

A third advantage is that the holographic recording and construction is insensitive to the exact position of the reference or reading beam on the hologram. The hologram can be moved but the focused spots remain stationary. This means that holographic memory systems are easier to align and less subject to problems of vibration than other optical memories.

Finally it is possible to record several holograms in different planes of one thick, sensitive material provided that different reference beam directions are used to record each hologram. The appropriate hologram is then read with a readout beam aligned at precisely the same angle as the original reference beam in the recording process.

Despite these advantages, the availability of suitable storage media has to date limited the development of holographic memories. Although photographic emulsion is satisfactory, it provides a permanent store only, which cannot be easily updated. Ideally the storage medium should have low write energy and high readout efficiency, be stable, but alterable so that data can be erased and new data written in. Potentially useful read/write/erase optically sensitive materials include thermoplastic photoconducting layers, magneto-optic MnBi and electrophotochromic KCl. Among the most attractive classes of recording materials are photorefractives (ref. 6.17), such as iron-doped lithium niobate ($\text{LiNbO}_3:\text{Fe}$), in which the irradiance of the incident light modulates the refractive index. A test system (ref. 6.18) using $\text{LiNbO}_3:\text{Fe}$ has stored 450 pages of 320×240 pixels with effectively a zero bit-error rate. The storage density approaches 10^8 bits cm^{-3} .

6.8 High energy applications of lasers

For several years now lasers have been used to perform a large number of materials processing functions such as cutting, drilling, welding, marking (or scribing) and the surface treatment of a wide range of materials. These include hard materials, for example diamond and ceramics, metals, wood, soft and pliable materials, and biological tissues. These processes have, of course, often been successfully undertaken without the need to resort to relatively advanced and expensive laser technology. Lasers, however, not only have significant advantages over conventional methods even for quite standard applications, but also may enable tasks to be undertaken which conventional techniques cannot accomplish.

The advantages of lasers, when compared with more conventional techniques for materials processing, include:

1. Laser radiation is a very 'clean' form of energy, in that no contaminating impurities need come into contact with the workpiece. In fact the working atmosphere can often be controlled to suit a particular task. Furthermore the radiation can be directed through a window in a vacuum chamber to provide the ultimate in contamination-free processing.
2. As we have seen laser beams, because of their high spatial coherence, may be focused into very small areas. Thus intense local heating can take place with limited effect on neighbouring areas.

3. It is comparatively easy to control the beam irradiances, and hence the energy delivered to the workpiece.
4. The beam can be readily directed into relatively inaccessible locations, and steered round sharp corners.
5. Most of the laser energy is deposited very near the surface of the workpiece (though there are exceptions – see section 2.8.1), thus enabling shallow regions to be treated without necessarily affecting the bulk.

It is probably true to say that the two lasers most commonly used for materials processing are the CO_2 and Nd:YAG lasers. This situation will perhaps change in the future, however, as other lasers become available, which are less expensive and offer different characteristics such as emission wavelength and output power. Comparing Nd:YAG and CO_2 lasers, the latter is the more versatile and is commercially available with a wide range of power outputs, up to several tens of kilowatts, at reasonable cost. There are some applications for which the Nd:YAG laser has advantages because of its shorter emission wavelength. For the same reason excimer lasers are increasingly being used in semiconductor device processing, where the shorter, blue/ultraviolet wavelengths enable smaller device features to be delineated via photolithography.

One of the most important aspects of laser processing is that of beam delivery to the workpiece. The success of a laser operation will often depend on the ability of the beam delivery system precisely to position a focused spot of radiation on the workpiece. In turn, the size of the spot, which is often of crucial importance, depends on the beam quality in terms of the spatial power density distribution, stability and number of oscillating modes.

Figure 6.24 shows a basic beam delivery system. It is, of course, not always possible to place the workpiece on a horizontal or other well-defined surface; indeed for many applications the beam must be steered into rather inaccessible locations. Until recently beam

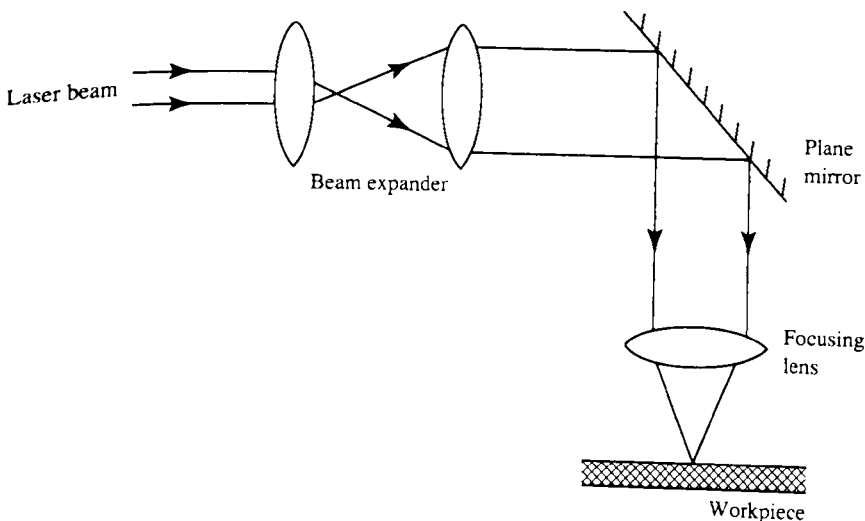


FIG. 6.24 Schematic layout of a basic laser beam delivery system.

steering was accomplished by articulated arms such as that illustrated in Fig. 6.25. Such systems, for CO₂ lasers for example, include a set of gold mirrors for reflection through as many as seven articulated joints. Though such systems are resistant to damage when used with high power lasers there are obvious limitations, including the separation of the laser and workpiece, arm flexibility and focusing accuracy, and the detrimental impact of the optical components on beam quality.

Increasingly, beam delivery systems use optical fibers, which, though of somewhat greater diameter than those described in Chapter 8, behave in essentially the same way. In practice the maximum diameter for silica fibers is 600 μm as beyond that the fiber is rather inflexible. While silica-based fibers similar to those used for fiber optic communications can be used for Nd:YAG and other near-infrared and visible output lasers, they cannot be used for CO₂ lasers where the emission wavelength of 10.6 μm is completely absorbed. Significant effort has been expended on the development of suitable fibers for transmitting laser radiation in the range 2–10 μm . Several materials have been developed for use, including zirconium fluoride, sapphire, germanium, chalcogenide glasses and silver halide as well as hollow waveguides. Of these chalcogenide glasses are quite effective for CO₂ lasers, while zirconium fluoride is preferred for near-infrared radiation.

In addition to improved flexibility and increased laser to workpiece separation, the use of fibers often leads to an improvement in the beam quality as it is delivered to the workpiece and consequently to more efficient processing (ref. 6.19). Laser beams, which may be multimode and have an asymmetric cross-section tend to be smoothed out as they propagate along the fiber thereby enabling a smaller focused spot to be obtained than otherwise would be the case. In fact the use of phase conjugate mirrors (PCMs) in non-linear optical phase conjugate (NOPC) systems leads to an output from the delivery fiber with a very high beam quality (ref. 6.20).

Another aspect of fiber delivery systems which has particularly revolutionised Nd:YAG, and other near-infrared, laser processing is the ability simultaneously to view the workpiece during processing. Many applications require very careful control of the focus position on

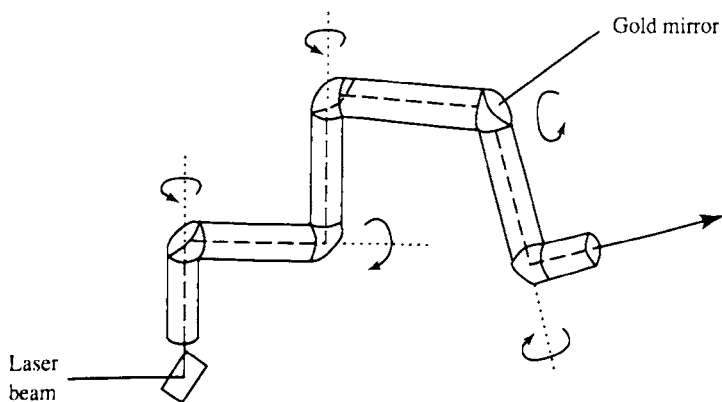


FIG. 6.25 Schematic diagram of an articulated-arm beam delivery system for use with CO₂ lasers in surgery – as many as seven joints may be included in the arm.

the workpiece, which may require manipulation of the beam by an operator or machine vision system. The only way of guaranteeing the required accuracy is to use the beam delivery optics to view the workpiece. This can be achieved for visible and near-infrared wavelengths, but not for CO₂ laser radiation, where the optics are either opaque to visible light or give a focus at a significant distance from that of the 10.6 μm radiation. Finally the use of fibers enables the radiation from a single laser to be split and shared between several workstations.

6.8.1 Industrial applications

A detailed theoretical analysis of the interaction between a laser beam and a material surface is extremely difficult to carry out. The material parameters which must be taken into account include the reflectance of the surface, thermal conductivity, specific heat capacity, latent heats of fusion and vaporization, melting and boiling points, and absorption of the laser radiation by the material vapour. Nevertheless the theory enables orders of magnitude calculations to be undertaken to estimate quantities such as the temperature rise and rate of removal of material for given laser power inputs. The calculations need to take into account whether the laser is operating in CW or in pulse mode, in which case the pulse repetition frequency is clearly important (ref. 6.21).

Given the limitations of space we shall consider, rather briefly, only three industrial applications, namely welding, cutting and drilling (or hole piercing); others are described in ref. 6.21.

In the basic *welding* process two metals, which may be the same or dissimilar, are placed in contact, and the region around the contact heated until the materials melt and fuse together. Careful control is required to ensure that sufficient heat is supplied to melt a sufficient volume of material, but not enough to give rise to significant vaporization of the material, which can lead to weak, porous welds. One of the problems is that the reflectance of most metals decreases dramatically as the temperature approaches the melting point, requiring further careful control of the incident energy.

Laser welding has to compete with many well-established techniques such as soldering, arc welding, resistance welding and electron beam welding. Laser welding has, however, a number of advantages, including:

1. Minimum heat input, which results in very little distortion of the workpiece.
2. Heating is localized and cooling is rapid so the neighbouring heat-affected zone is small.
3. There is no physical contact with external components.
4. Dissimilar materials can be welded, which is often difficult using other techniques.
5. The process can be easily automated.
6. Faster weld rates can be obtained than from other techniques.
7. The high quality of the resulting welds.

Welding is normally carried out using a shielding (as shown in Fig. 6.26) inert gas such as argon or helium to cover the weld area to prevent oxidation of the metals, which results in poor welds. The gas also helps to remove any metal vapour which may be created and which

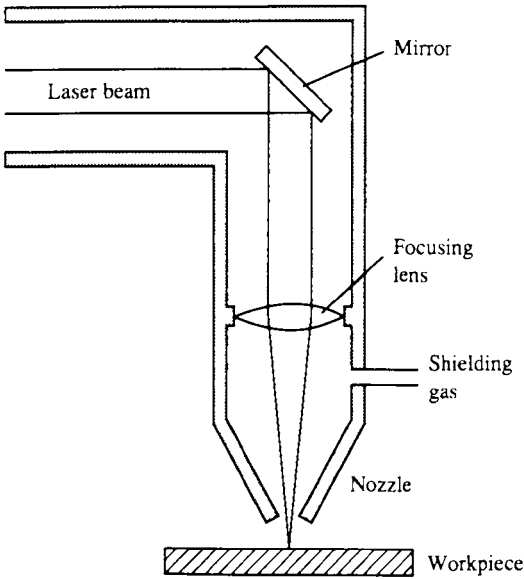


FIG. 6.26 Schematic beam focusing head design for laser welding when using a shielding gas.

may be deposited on the focusing lens. Metal vapours can also be highly absorbant, and in some cases can prevent most of the laser radiation from reaching the workpiece.

Both CW and pulse lasers are used for welding. For example, continuous seam welds are performed almost exclusively with CW lasers operating at 500 W or greater. While CW lasers produce continuous welds, pulse lasers produce a train of spot welds as the beam moves over the workpiece. The separation of the spot welds depends, of course, on the pulse repetition frequency and scan speed so that they may overlap and produce a quasi-continuous weld. Clearly in situations where only a single or small number of spot welds are required, a pulse laser is entirely appropriate, in which case Nd:YAG or Nd:glass lasers are often used.

When using multikilowatt CO_2 or high energy pulse mode lasers the welding process becomes rather more complicated. In these situations, when the laser beam first strikes the surface a significant amount of material may be vaporized forming a small hole known as a *keyhole* (Fig. 6.27). Laser energy which subsequently enters this hole is trapped and is carried deeper into the material than otherwise would be the case. The material around the keyhole then melts and fills in the hole and later solidifies to form the weld as the laser scans across the surface. Were it not for this phenomenon weld depths would be limited to only about a millimetre or so. With keyholing, however, weld depths of several tens of millimetres have been achieved with high power CO_2 lasers, thereby facilitating the fabrication of large structures.

At the other end of the scale lasers are also used in microelectronics to weld minute electrical contacts, and to weld small devices accurately into place. Again the advantages of being able to deposit exactly the right amount of heat precisely to the required areas are extremely important.

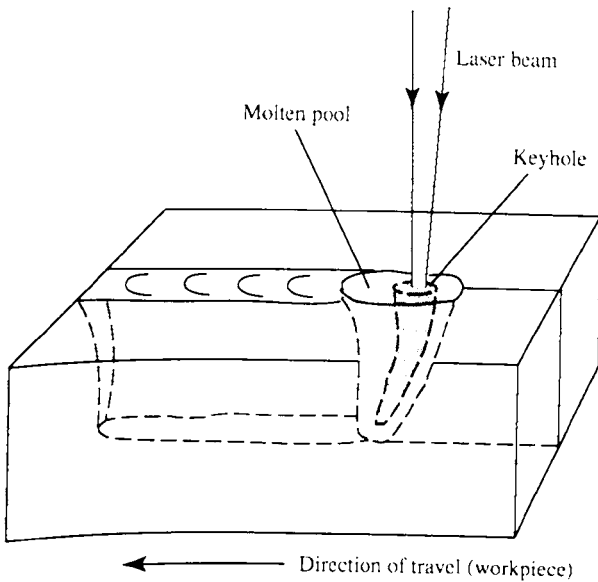


FIG. 6.27 Formation of a 'keyhole' during high power laser welding.

Industrial laser *cutting* uses very much the same range of lasers as is used in welding, and the advantages of laser cutting over conventional techniques are similar to those of welding, and essentially independent of the type of laser used. In cutting the aim is to vaporize the material as quickly as possible to produce a heat-affected zone (HAZ) which is as narrow as possible. Laser cutting is a gas-assisted process in which a gas, under pressure, forces molten material from the region of the cut or *kerf*. In the case of reactive materials such as metals oxygen is normally used; this enhances the cutting rate and quality of the cut edges, which are left with a thin oxide coating. In some circumstances, for example if the metal edge is to be subsequently welded, the oxide layer is undesirable and an inert gas must be used. Similarly when non-metallic materials such as ceramic, wood, paper or plastics are to be cut, oxidation is better avoided and again an inert gas is employed.

The gas stream, as in welding, is usually delivered co-axially with the laser beam, and it is important to maintain control over the pressure and flow rate during cutting. In many cases carefully shaped nozzles have been designed to optimize the cutting process.

Thicknesses of up to about 5 mm of steel can be readily cut with 500 W, CO_2 lasers, and even greater thicknesses can be dealt with using higher power lasers. The faces of the cut, or *kerf*, in comparatively thick metals and non-metals are often surprisingly straight sided. This results from a light-guiding effect due to multiple reflections of the laser beam from the sides of the *kerf* as illustrated in Fig. 6.28. The amount of remaining taper can be controlled to a certain extent by the positioning of the focal point of the delivery lens. In metal cutting it is best to focus on the surface of the workpiece, while for non-metals better results are obtained if the focus is below the surface.

Laser cutting systems are frequently essentially numerically controlled systems or robots. In the case of cutting sheet metal, for example, a two-axis numerical controller is adequate.

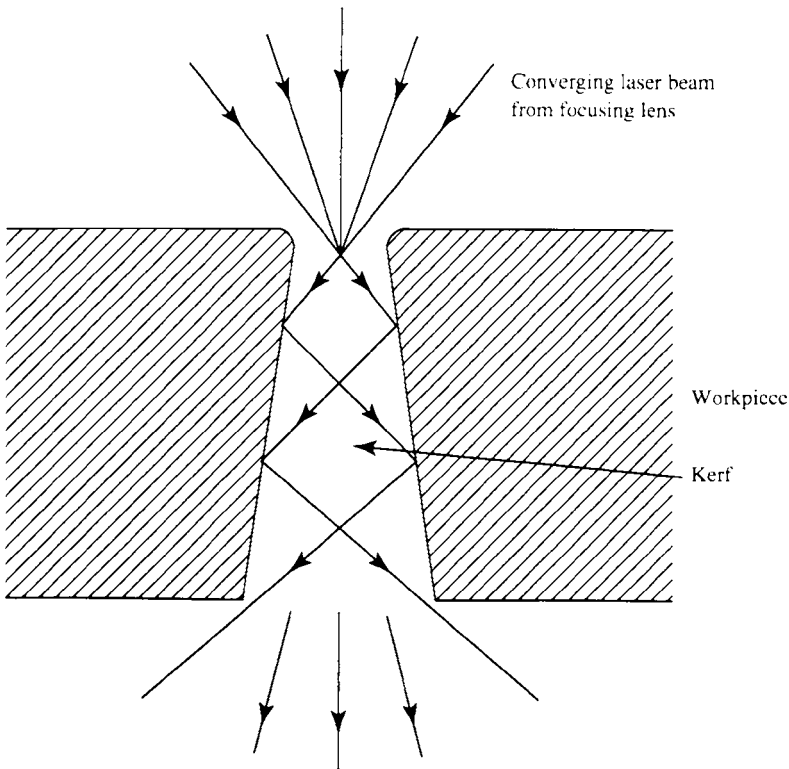


FIG. 6.28 Sketch of the kerf produced by laser cutting. The laser beam is channelled through the material by multiple reflections from the walls producing a waveguiding effect. A similar effect occurs in laser drilling.

In more complicated processing, however, four- or five-axis numerical control with automatic location sensing and feedback control of the focused beam may be required. The flexibility of such systems is one of the main advantages of laser cutting. For example, where several 'options' in a processed part are required, this can be achieved by appropriate programming of the system, without the need for frequent tooling changes. Such advantages have led to the increasing use of laser processing systems in the automotive industry.

Scribing and *marking* are similar to cutting, though the depth of the cut is very much reduced. In scribing a line of weakness is created in the surface of a material prior to breaking it along the line. This technique is widely used in the semiconductor industry where, for example, processed silicon wafers are scribed prior to breaking them into individual chips. Similarly laser marking is a convenient way of engraving identification numbers, symbols or logos onto a variety of material objects.

Hole piercing or *drilling* was one of the first industrial uses of lasers, when in 1965 a diamond die for wire drawing was drilled with a ruby laser in about one-hundredth of the time taken by conventional means. Since then lasers have been used to drill an amazingly wide range of materials including cigarette filter paper, plastic and rubber baby bottle nipples,

aerosol can nozzles, ceramics and glasses, and titanium alloy parts used in turbine blades. Again the basic advantages of lasers for drilling are essentially the same as for welding and cutting. It is a technique in which highly reproducible holes can be made at extremely high rates, frequently in materials which are otherwise difficult to machine, without any tool contact.

Hole dimensions are described in terms of the *aspect ratio*, that is the ratio of the hole depth to hole diameter, and *taper*, which is the ratio of the entrance diameter to the mid-section diameter. As with cutting the sides of the hole are straighter than we may have anticipated because of the waveguiding effects of the sides of the hole (again refer to Fig. 6.28). This effect is more apparent in glass-like materials, where aspect ratios of 25 may be achieved, whereas in metals the figure is nearer 12. As might be expected, it is found that higher aspect ratios are favoured by slowly converging laser beams.

In most drilling operations the surface temperature of the target material is raised above the boiling point of the material. A little thought reveals that, for a given amount of heat energy delivered to the surface, the shorter the pulse duration the better. During longer pulses the heat deposited has more time to diffuse into the material, resulting in a relatively larger volume of material being heated to a lower temperature. We therefore expect pulse lasers to be more effective than CW ones, assuming comparable average powers. For hole drilling in metals the CO₂ laser suffers from the problem of poor initial energy absorption, and Nd:YAG lasers are preferred. For non-metals the situation is reversed. Excimer lasers are effective in drilling very clean holes in polymeric and other organic materials because of a non-thermal interaction of the ultraviolet radiation produced with the material.

6.8.2 Medical applications

One of the most rapidly developing areas of laser applications is perhaps in medicine. Many medical applications arise directly from the laser's materials processing capabilities, described in the previous section. Similarly many of the advantages of lasers over conventional techniques arise from the precision and controllability, which minimize disruption and damage to the patient, and the possibility of delivering the laser radiation to somewhat inaccessible regions of the body. Again, because of the limitations of space, we shall confine ourselves to brief descriptions of a small number of representative applications—for others see ref. 6.22.

One of the first, and still one of the commonest, applications is the use of CO₂ lasers in general surgery. The 10.6 μm emission wavelength is strongly absorbed by water molecules present in tissue, and the subsequent evaporation of the water leads to the physical removal of the tissue (the resulting debris is usually removed by a vacuum pump). The advantages over physical cutting include the limited damage to adjacent tissue and the cauterizing effect of the radiation on the blood vessels, which reduces bleeding. A recently introduced alternative to CO₂ lasers for surgery and related applications is the erbium laser, that is Er:YAG. The erbium laser emits at a wavelength of 2.94 μm , which is also strongly absorbed by water molecules, and the laser is consequently particularly useful for precise, localized tissue removal or ablation. Commercially available optical fibers transmit radiation more efficiently at 2.94 μm than at 10.6 μm , and it is likely that zirconium fluoride fibers will replace

articulated arms at this latter wavelength for many applications in the near future (ref. 6.23). Erbium lasers are also beginning to compete with other lasers in dentistry and especially ophthalmology, which is an area where lasers have a number of applications.

One of the earliest applications in ophthalmology was in the treatment of detached retinas using light from a ruby laser, which with a wavelength of 694 nm readily passes through the cornea and other transparent regions, to be absorbed by the red blood cells at the back of the eye. A pulse of radiation causes a lesion and when this heals the scar tissue formed reattaches the retina (in effect a spot weld). Similarly laser radiation absorbed at the back of the eye can be used to treat degenerative conditions associated with diabetes, which can lead to blindness.

The hole drilling capability of lasers has been used to drill small holes (or fistulas) of about 300 μm diameter through the sclera to reduce the increased intraocular pressure in the eye resulting from glaucoma. Finally, another application is what is often referred to as *corneal sculpting* or *photorefractive keratectomy* (PRK), in which the shape of the cornea is modified by removing (ablating) thin layers, some 50 μm in thickness, from appropriate regions. This changes the curvature of the cornea so that corrections for myopia (near-sightedness) and hypermetropia (far-sightedness) can be effected. ArF excimer lasers ($\lambda = 193 \text{ nm}$) are probably most frequently used for these increasingly common operations, as the emission wavelength is very effective in ablating corneal tissue. However, high frequency harmonics of various solid lasers have also been used.

Lasers can also be used in the treatment of cancer, particularly in otherwise inaccessible parts of the body, such as the larynx and neck of the womb. Indeed laser radiation has been used very successfully for several years in the treatment of the early stages of cervical cancer. A particularly promising technique for the treatment of tumours is *photodynamic therapy*. In this treatment, the patient is injected with a specially designed dye substance called HpD which accumulates in cancerous tissue, but which is rejected by healthy tissue. When exposed to laser radiation at a wavelength of about 630 nm, the HpD undergoes a series of photochemical reactions resulting in the formation of a chemical which kills the cancerous tissue.

Lastly, in some applications the laser radiation is transmitted down an extremely fine optical fiber, which can be introduced into arteries, for example, using catheters. Radiation can then be delivered to remote parts of the body as in the removal of deposits of plaque, a fatty material, which reduces the flow of blood along the coronary artery. To ensure accurate positioning of the laser beam, and for example to prevent accidentally creating a hole in the artery, a second fiber comprising a coherent fiber bundle (section 10.2.1) or endoscope is included so that the surgeon has a view of the process, as illustrated in Fig. 6.29. Similar techniques can be used to remove other obstructions such as blood clots.

With the majority of laser treatments the side effects are minimal compared with conventional techniques, and patients suffer significantly less discomfort, and can often be treated as outpatients, which is a major factor in terms of cost and convenience.

6.8.3 Laser-induced nuclear fusion

For many years, research has been directed towards a system for producing controlled thermonuclear reactions to generate energy. Nuclear fusion of light elements occurs within a

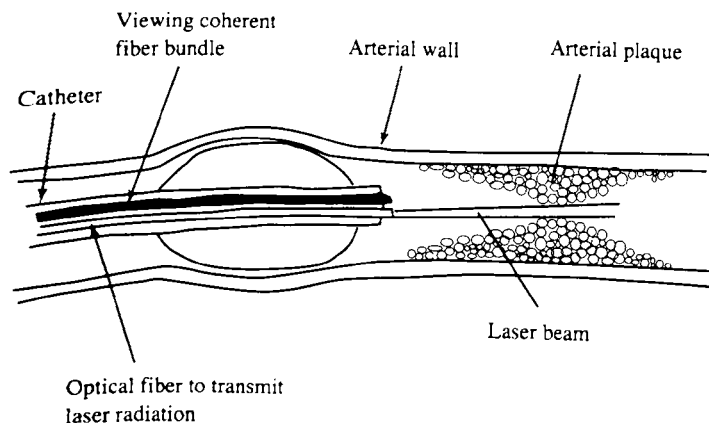


FIG. 6.29 Removal of arterial plaque using laser radiation transmitted down a fine optical fiber inserted into the artery. A viewing fiber bundle is also included in the 'tube' inserted into the artery.

very high temperature plasma such as exists in the sun. Until recently, laboratory experiments aimed at reproducing these conditions were based entirely on magnetic confinement of the plasma, and these have advanced considerably in recent years using the tokamak concept. Since the early 1970s, however, with the advent of very high power lasers an alternative way of producing suitable conditions has been under investigation, namely inertial confinement fusion.

The basic concept simply involves focusing very high power laser radiation onto a target. In the early research the target comprised glass pellets approximately $50\text{ }\mu\text{m}$ in diameter containing a mixture of deuterium and tritium gases at high pressures or pellets of frozen heavy water (D_2O) and extra heavy water (T_2O). More recently, however, carbon-based plastic microspheres with a $500\text{ }\mu\text{m}$ diameter and extremely thin walls ($2 \times 10^{-3}\text{ }\mu\text{m}$) have been fabricated to hold the gas mixture (ref. 6.24). A number of laser beams are directed onto the target simultaneously from symmetrically arrayed directions. Absorption of the laser radiation at the surface of the pellet causes ablation (burning off) of the outer material and an implosion of the contents. The implosion is caused by a compressional wave driving radially into the material from the periphery thereby squeezing the pellet into a very dense core. Very high temperatures, in excess of 10^8 K , are produced within this core and at these temperatures the velocities of the deuterium and tritium atoms are so great that the electrostatic repulsion of the positive nuclei is overcome and the atoms undergo fusion. A typical fusion reaction is



The reaction yields a helium atom and an energetic neutron. For many such reactions to take place within the compressed pellet, the high temperature must be maintained for about 1 ps and the compression must be in the region of $10^4:1$. These conditions require enormous laser pulse energies.

While very large numbers of neutrons have been observed following a laser pulse, these

are still considerably lower than that required to reach 'scientific breakeven', which is defined as the level at which the thermonuclear energy generated equals the laser energy input. Calculations indicate that a laser input pulse of perhaps 10^{14} W with a subnanosecond duration may be necessary to achieve breakeven and also that laser-induced fusion may be most efficient in the wavelength range 300–600 nm. A series of lasers with steadily increasing capability has been built over the last few years for this research. These include the Shiva Nd:glass laser at the Lawrence Livermore Laboratory, California, which has 20 amplifier chains each delivering about 1 TW in a 100 ps pulse to the fuel pellets. The individual amplifier chains may have quite contorted geometries to equalize their lengths to ensure that they all deliver their energy to the target at precisely the same instant. The energy supplied to the amplifier chains is derived from a single Nd:glass master laser. A similar system called Nova, which is 20 times more powerful, was completed in 1985. Nova comprises eight separate Nd:glass amplifier beam lines, all fed by a single oscillator. Each beam line has an input energy per pulse of approximately 100 nJ, and a total gain of 10^{11} gives output pulses of some 10 kJ, thereby producing a combined energy of 80 kJ with a pulse duration of 1–2 ns, so that the peak power approaches 10^{14} W. The 1.06 μm emission wavelength can be frequency doubled to 525 nm to increase the efficiency of the fusion process.

An even more powerful laser facility, The National Ignition Facility, has just been approved by the US Congress (ref. 6.25) with a target completion date of 2002. The facility, which again is based on Nd:glass amplifiers, will have 192 separate beam lines in which the emission from a single laser will be amplified and undergo two levels of frequency conversion (from a wavelength of 1.06 μm to 525 nm to 350 nm) prior to hitting the target. Each beam line will produce output energies of 15 kJ in pulses of 3 ns duration, so the total design energy is about 3 MJ and peak powers will approach 10^{15} W. Although neodymium-based lasers and amplifiers are attractive for fusion studies in view of the enormous powers they can generate, it is debatable as to whether their use in practical power generation will be feasible. The main reason for this is that they cannot be cooled sufficiently rapidly to allow the repetition rate of 100 pulses per second which may be required. Gas lasers, however, can be cooled quickly by convection and the use of CO_2 lasers has been investigated. As suggested above, however, the CO_2 emission wavelength may be less favourable than shorter wavelengths, and for that reason there has been some interest in the use of excimer lasers, which can emit in the range 200–600 nm range, to initiate fusion reactions. Developments within the next few years may yet realize the goal of fusion-based power generation (ref. 6.26).

PROBLEMS

- 6.1 If the halfwidth of the He–Ne 632.8 nm transition is 1500 MHz, what must be the length of the laser cavity to ensure that only one longitudinal mode oscillates? Estimate the accuracy to which the temperature must be controlled if the frequency stability is to be better than 10^8 Hz. Take the coefficient of expansion of the laser tube to be $10^{-6}^\circ\text{C}^{-1}$.
- 6.2 If a cavity mode burns holes in the gain versus velocity curve for the 632.8 nm He–Ne transition at the spectral halfwidth points (halfwidth = 1500 MHz), what is the velocity of the atoms involved in the hole burning?

- 6.3 If one of the mirrors in the Michelson interferometer moves with velocity v , show that the rate at which the fringes cross the field of view is $2v/\lambda$. Show that this result can be obtained by calculating the frequency of the beats generated between the light reflected from the stationary mirror and the moving mirror. (The Doppler shift of the frequency of light reflected from moving objects is the basis of many laser velocimeters – see ref. 6.4e, pp. 306–15 and 334.)
- 6.4 Calculate the Doppler frequency shift for light ($\lambda_0 = 500$ nm) reflected from an object moving at 20 m s^{-1} ; what implications does your answer have for the frequency stabilization of lasers used in Doppler velocimeters?
- 6.5 Explain how a laser may be used to measure the width of a narrow slit from the Fraunhofer diffraction pattern produced by the slit. The second minimum in the diffraction pattern, which is formed in the focal plane of a lens of focal length 0.5 m, is 2 mm from the central maximum. What is the width of the slit (take $\lambda_0 = 632.8$ nm)? Discuss how this technique may be extended to the measurement of the diameter of a thin wire and hence used in controlling the diameter of the wire during production.
- 6.6 A CW argon laser emits 1 W at $\lambda_0 = 488$ nm; if the beam divergence is 0.5 mrad and the diameter of the beam at the output mirror is 2 mm, calculate the brightness of the laser. To what *photometric* brightness (or luminance) does this correspond?
- 6.7 Compare the mode-locked pulse width, and the separation between pulses produced by:
- (a) an He–Ne laser operating at 632.8 nm, with a mirror spacing of 0.5 m, given that the emission linewidth is 1.5×10^9 Hz;
 - (b) an Rh 6G dye laser operating over its full gain bandwidth of 570–640 nm, with a mirror spacing of 2.0 m;
 - (c) an Nd:glass laser in which mirrors are formed on the ends of a laser rod 0.2 m long, and assuming that there are 3000 participating longitudinal modes; take the refractive index of Nd:glass to be 1.54.
- 6.8 If the halfwidth of the $10.6 \mu\text{m}$ transition of a low pressure CO_2 laser is 60 MHz, calculate the coherence length of the laser. If the cavity length is 2 m, show that not more than one mode will oscillate. If we take the width of the Fabry–Perot resonances as an (over)estimate of the spectral width of the mode, calculate the coherence length – take the mirror reflectance R to be 0.95. (See ref. 1.1c, pp. 90–6, for example.)
- 6.9 What is the total energy release in the nuclear reaction given in eq. (6.15)? Why does this value differ from that given in the text for the energy of the neutron? (You will need a table of nuclear masses.)

REFERENCES

- 6.1 (a) G. M. S. Joynes and R. B. Wiseman, 'Techniques for single mode selection and stabilisation in helium-neon lasers', in H. G. Jerrard (ed.) *Electro-optics/Laser International '80 UK* (Conference Proceedings), IPC Science & Technology Press, London 1980, p. 163.

- (b) G. E. Moss, 'High power single-mode HeNe laser', *Appl. Opt.*, **10**, 2565, 1971.
- (c) J. Hawkes and I. Latimer, *Lasers: Theory and Practice*, Prentice Hall International, Hemel Hempstead, 1994, Section 6.6.3.
- 6.2 Ref. 6.1c, Section 6.6.4.
- 6.3 E. E. Basch (ed.), *Optical Fibre Transmission*, Howard W. Sams, Indianapolis, 1987, Chapter 9, Section 4.
- 6.4 H. A. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, Englewood Cliffs, NJ, 1984, Sections 8.1 and 8.2.
- 6.5 (a) N. K. Dutta *et al.*, 'Single longitudinal mode operation of a semiconductor laser using a metal film reflection filter', *IEEE J. Quantum Electron.*, **QE-21**, 559, 1985.
- (b) W. T. Tsang, 'The cleaved coupled cavity laser', in *Semiconductors and Semimetals*, Vol. 22B, *Semiconductor Lasers I*, W. T. Tsang (ed.), Academic Press, New York, 1985, p. 263.
- 6.6 A. Yariv, *Optical Electronics* (4th edn), Holt-Saunders, New York, 1991, Section 6.7.
- 6.7 (a) J. Wilson and J. F. D. Hawkes, *Lasers: Principles and Applications*, Prentice Hall, Hemel Hempstead, 1987.
- (b) M. Ross (ed.), *Laser Applications*, Vol. 1, Academic Press, New York, 1972.
- (c) S. S. Charschan (ed.), *Lasers in Industry*, Van Nostrand Reinhold, New York, 1972.
- (d) J. E. Harry, *Industrial Lasers and their Applications*, McGraw-Hill, Maidenhead, 1974.
- (e) J. F. Ready, *Industrial Applications of Lasers*, Academic Press, New York, 1978.
- (f) *Laser Focus*, Advanced Technology Publications, Newton, MA (a controlled circulation publication available to those working in laser-related fields).
- (g) *Electro-Optical Systems Design*, Kiver Publications, Chicago (a controlled circulation publication).
- (h) *Photonics Spectra*, Laurins Publishing Company, Pittsfield, MA.
- (i) *Lasers and Applications*, Laurins Publishing Company, Pittsfield, MA.
- 6.8 (a) M. V. Klein and T. E. Furtak, *Optics* (2nd edn), John Wiley, New York, 1986, Chapter 8.
- (b) See ref. 6.1c, Section 2.4.
- 6.9 M. Anson, 'Laser speckle vibrometry: a technique for analysis of small vibrations', *Proc. Physiol. Soc. (London)*, **300**, 8P, 1980.
- 6.10 (a) A. Yariv, *op. cit.*, Chapter 5.
- (b) J. E. Harry, *op. cit.*, pp. 111–40.
- (c) J. F. Ready, *op. cit.*, Chapters 13–16.
- 6.11 (a) See ref. 6.7a, Section 7.3.
- (b) G. Bouwhuis *et al.*, *Principles of optical disc systems*, Adam Hilger, Bristol, 1985.
- (c) Jordan Isailovic, *Videodisc and Optical Memory Systems*, Prentice Hall, Englewood Cliffs, NJ, 1985.
- (d) T. V. Higgins, 'Optical storage lights the multimedia future', *Laser Focus World*, Sept., 103, 1995.
- 6.12 (a) R. N. Zare, 'Laser separation of isotopes', *Sci. Am.*, Feb., 1977.

- (b) M. B. Radunsky, 'How to extend frequency ranges of tunable lasers', *Laser Focus World*, July, 77, 1996.
- (c) B. J. Orr *et al.*, 'Spectroscopic applications of pulsed tunable optical parametric amplifiers', in F. J. Duarte (ed.) *Tunable Laser Applications*, Marcel Dekker, New York, 1995.
- (d) See ref. 6.7a, Section 4.2.
- 6.13** (a) A. Yariv, *op. cit.*, Chapter 4.
- (b) J. F. Ready, *op. cit.*, Chapter 11.
- 6.14** D. Gabor, 'A new microscopic principle', *Nature*, **4098**, 777, 1948.
- 6.15** (a) G. W. Stroke, *An Introduction to Coherent Optics and Holography* (2nd edn), Academic Press, New York, 1969.
- (b) R. H. Collier, C. B. Burkhart and L. H. Lin, *Optical Holography*, Academic Press, New York, 1972.
- (c) H. M. Smith, *Principles of Holography*, Wiley-Interscience, New York, 1975.
- 6.16** (a) Yu I. Ostrovsky, *Holography and its Applications*, Mir, Moscow, 1977.
- (b) See ref. 4.7a, Chapter 6.
- 6.17** F. H. Mok *et al.*, *Opt. Lett.*, **21**, 816, 1996.
- 6.18** (a) J. Ashley *et al.*, 'Holographic storage promises high data density', *Laser Focus World*, Nov., 51, 1996.
- (b) J. H. Hong and D. Psaltis, 'Dense holographic storage promises fast access', *Laser Focus World*, Apr., 119, 1996.
- (c) A. Strass, 'Holographic memories target terabyte storage', *Opto and Laser Europe*, Sept., 31, 1996.
- 6.19** T. R. Kugler, 'Fibre delivers the goods, for Nd:YAG applications', *Photonics Spectra*, Sept., 103, 1996.
- 6.20** (a) D. A. Rookwell *et al.*, 'Turning up the power in fibre/laser systems', *Photonics Spectra*, Sept., 103, 1996.
- (b) See ref. 6.7a, Section 3.10.
- 6.21** (a) *Ibid.*, Chapter 5.
- (b) J. T. Luxon and O. E. Parker, *Industrial Lasers and their Applications*, Prentice Hall, Englewood Cliffs, NJ, 1992.
- 6.22** (a) A. F. Carruth and A. L. McKenzie, *Medical Laser Science and Clinical Practice*, Adam Hilger, Bristol, 1986.
- (b) *Biophotonics International*, a Laurin publication.
- (c) *Lasers in Surgery and Medicine*.
- (d) *Ophthalmology*.
- (e) *Opt. Lett.*
- 6.23** K. Vogler and M. Reindl, 'Improved erbium laser parameters for new medical applications', *Biophotonics Int.*, Nov., 40, 1996.
- 6.24** R. F. Service, 'Small spheres lead to big ideas', *Science*, **267**, 327, 1995.

- 6.25 (a) V. Kiernan, 'Laser industry gets boost from last-minute FY 1997 appropriations', *Laser Focus World*, Nov., 71, 1996.
- (b) W. J. Hogan *et al.*, 'National Ignition Facility design focuses on optics', *Laser Focus World*, Nov., 107, 1996.
- 6.26 (a) C. Yamanaka, 'Prospect of laser fusion research', *Laser Phys.*, 6, 506, 1996.
- (b) R. R. Johnson, 'Laser fusion in the United States', *Opt. Photonic News*, 6(3), 16, 1995.
- (c) C. Yamanaka, 'Past, present and future of laser fusion research', *AIP Conference Proceedings*, No. 1369/PT1, 3, 1996.

Photodetectors

The optical detectors discussed in this chapter may be classified as either *thermal* or *photon* devices. In thermal detectors, the absorption of light raises the temperature of the device and this in turn results in changes in some temperature-dependent parameter (e.g. electrical conductivity). As a consequence, the output of thermal detectors is usually proportional to the amount of energy absorbed per unit time by the detector and, provided the absorption efficiency is the same at all wavelengths, is independent of the wavelength of the light. In photon detectors, on the other hand, the absorption process results directly in some specific quantum event (such as the photoelectric emission of electrons from a surface) which is then 'counted' by the detection system. Thus the output of photon detectors is governed by the rate of absorption of light quanta and not directly on their energy. Furthermore, all the photon processes considered here require a certain minimum photon energy to initiate them. Since the energy of a single photon is given by $E = h\nu = hc/\lambda_0$ (see eq. 1.2), photon detectors have a long wavelength 'cut-off', that is a maximum wavelength beyond which they do not operate.

A problem encountered with photon detectors operated in the infrared is that the photon energies involved become comparable with the average thermal energies ($\approx kT$) of atoms in the detector itself. A relatively large number of quantum events may then be generated by thermal excitation rather than by light absorption and will thus constitute a source of noise. The obvious way to reduce this noise signal is to reduce the temperature of the detector; indeed most photon detectors operating above a wavelength of $3\text{ }\mu\text{m}$ or so must be cooled to liquid nitrogen temperatures (77 K) or below.

7.1

Detector performance parameters

Before we embark on a detailed discussion of individual detectors it is useful to review some of the parameters that are commonly used to assess the performance of a detector. The *responsivity* R is defined as the ratio of the detector output to its input. The units used will depend both on the type of detector and its intended use, but typically will be amps (or volts) per watt. If the device is intended for use in the visible, then amps per lumen are sometimes used. The responsivity will vary with wavelength and should be designated R_λ ; the *spectral response* is usually given as a curve of R_λ versus wavelength.

It is obviously important that a detector is able to follow the variations in intensity of incoming radiation as closely as possible. For example, a sudden step change in irradiance should produce a similar step change in the output of the detector. Inevitably there are limits

to all detector performances in this respect. Many detectors respond to such a step change with an exponential rise or fall, characterized by a *response time* or *time constant* τ . As is shown in Appendix 5, this implies that the detector will have a responsivity which depends on the light modulation frequency f according to

$$R(f) = \frac{R(0)}{(1 + 4\pi^2 f^2 \tau^2)^{1/2}} \quad (7.1)$$

The upper operating frequency (the *cut-off frequency*, f_c) of a detector can be defined as the frequency at which the electrical power output falls to one-half of its maximum value. Since the responsivity of an optical detector inevitably involves either a current or voltage then we are looking for the frequency at which the responsivity falls to $1/\sqrt{2}$ of its maximum value. If the responsivity is given by eq. (7.1) where $R(f)$ has its maximum value at $f=0$, the cut-off frequency is given by $f_c = 1/(2\pi\tau)$. Obviously a detector can only give a faithful representation of an input signal if the highest frequency content of the signal is less than f_c .

Another important aspect of a detector is its *sensitivity*, that is its ability to measure very small optical signals. The limitations on the size of signal that can be detected come about because of the presence of *noise*; that is, an output from the detector that is completely uncorrelated with the signal. The ratio of the magnitude of the output due to the signal to that of the noise is referred to as the *signal-to-noise ratio* (S/N). It is generally assumed that it will be difficult to measure a signal when it generates an r.m.s output equal to that generated by the noise (i.e. when $S/N = 1$). An indication of the size of the minimum detectable signal (or sensitivity) is given by the *noise equivalent power* (NEP). This is defined as the power of sinusoidally modulated monochromatic radiation which would result in the same r.m.s. output signal in an 'ideal' noise-free detector. Statements of the NEP should be given along with a statement of the modulation frequency, detector bandwidth, detector temperature and detector area. It is useful to remember that many noise sources produce what is known as *white noise*, where the noise power within a bandwidth Δf is proportional to Δf . Because both current and voltage are proportional to the square root of electrical power, noise current and noise voltage are then proportional to $\Delta f^{1/2}$.

If we assume that the noise power generated in a detector is proportional to its sensitive area A , then the noise current (or voltage) will vary as $A^{1/2}$. Thus we may define a unit NEP* which takes into account the effects of variable bandwidth and detector area where

$$\text{NEP}^* = \frac{\text{NEP}}{(A\Delta f)^{1/2}}$$

In fact it is the reciprocal of this quantity, known as the *specific detectivity* (D^*), that is commonly used. Thus

$$D^* = \frac{(A\Delta f)^{1/2}}{\text{NEP}} \quad (7.2)$$

The value of D^* for a particular detector will depend on the wavelength of the signal radiation and the frequency at which it is modulated, and is often quoted in the format $D^*(\lambda_0, f)$.

Curves showing the variation of D^* with λ_0 for several of the detectors that will be

discussed in this chapter are shown in Fig. 7.1. Also indicated on this diagram is the theoretical D^* curve for a detector where the noise arises from fluctuations in carrier generation caused by the presence of blackbody radiation at 300 K. It will be seen that, at their best, a number of detectors approach this theoretical limit.

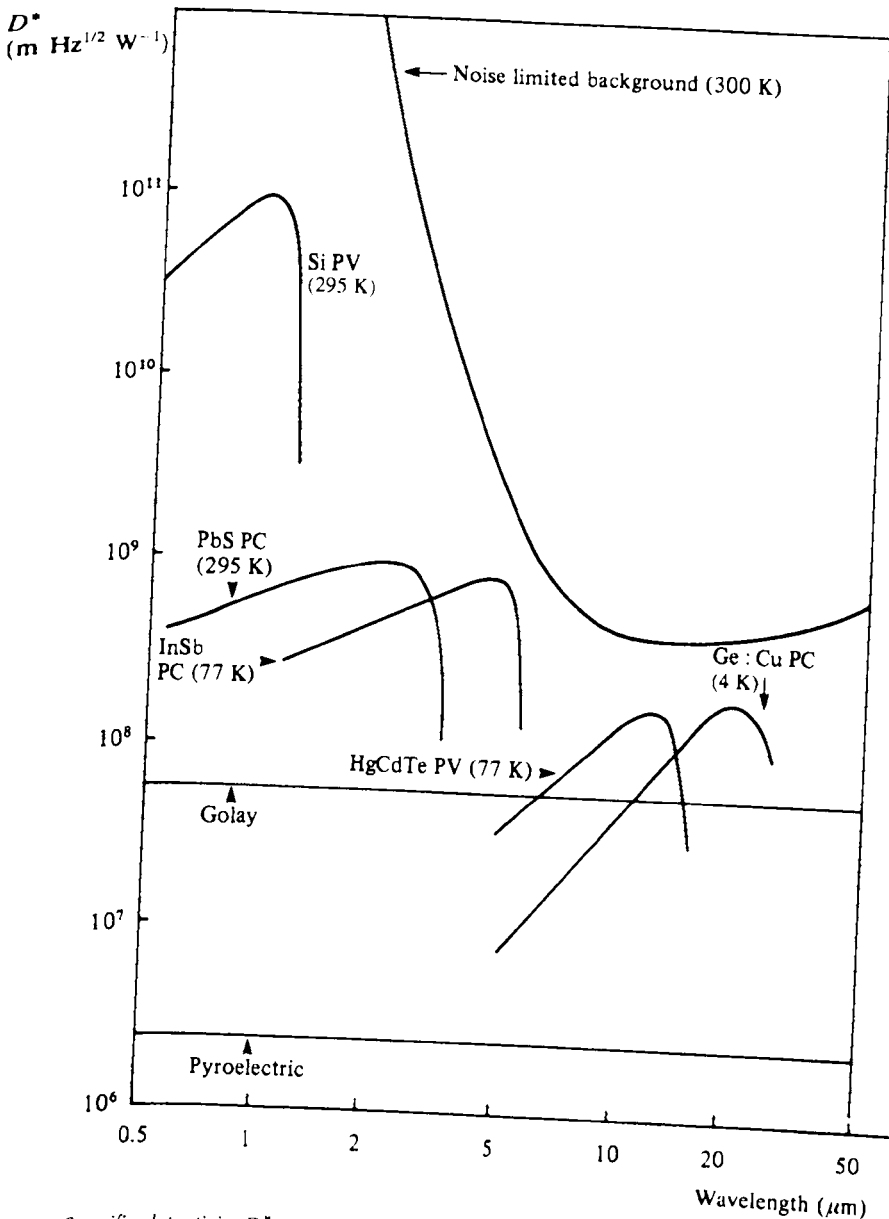


FIG. 7.1 Specific detectivity D^* as a function of wavelength for a number of representative photodetectors (PC, photoconductive; PV, photovoltaic).

Provided that other criteria such as frequency response are met, and that the noise is not background radiation limited, then at a particular wavelength the detector with the highest D^* value is generally the best choice.

EXAMPLE 7.1 Detector sensitivity

From Fig. 7.1 a PbS detector has a D^* of 10^9 m Hz^{1/2} W⁻¹ at a wavelength of 2 μ m. If we assume a detector area of 25×10^{-6} m² and a detection bandwidth of 10 kHz then the sensitivity of the device at this wavelength may be calculated from eq. (7.2). Thus

$$\begin{aligned} \text{NEP} &= \frac{(25 \times 10^{-6} \times 10^4)^{1/2}}{10^9} \text{ W} \\ &= 5 \times 10^{-10} \text{ W} \end{aligned}$$

7.2 Thermal detectors

To gain an insight into the performance characteristics of thermal detectors, we consider the behaviour of the simple model shown in Fig. 7.2. The incoming radiation is absorbed within the sensing element of heat capacity H . This is connected to a heat sink at constant temperature T_s via a heat conducting link which has a thermal conductance G . If the instantaneous rate of heat absorption is given by W , then during a small time interval δt the heat absorbed is $W\delta t$. If we let the temperature of the element be $T_s + \Delta T$ then during the same time interval the amount of heat lost through the thermal link is $G\Delta T\delta t$. The difference between these two represents the amount of heat available to raise the temperature of the

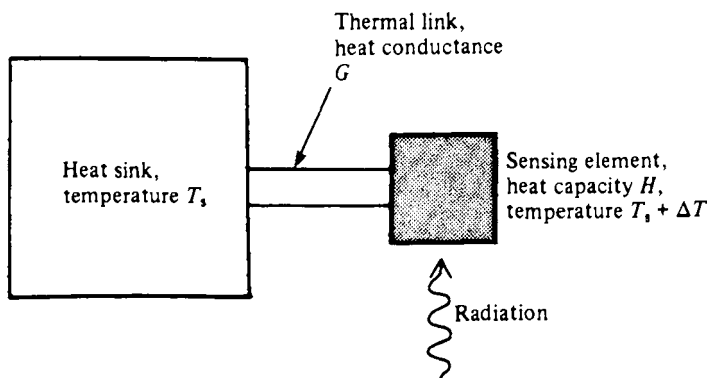


FIG. 7.2 Model of a thermal detector used to derive the frequency response characteristics. The incoming radiation causes the instantaneous temperature of the sensing element to be $T_s + \Delta T$. The element is connected via a conducting link (of conductance G) to a heat sink which remains at the temperature T_s .

element. Hence we may write

$$W\delta t - G\Delta T\delta t = H\delta(\Delta T)$$

If we take the limit $\delta t \rightarrow 0$ we obtain

$$W = H \frac{d(\Delta T)}{dt} + G\Delta T \quad (7.3)$$

Now suppose that W has a time dependence given by $W = W_0 + W_f \cos(2\pi ft)$ where $W_0 \gg W_f$ and also that ΔT can similarly be written $\Delta T = \Delta T_0 + \Delta T_f \cos(2\pi ft + \phi_f)$. By substituting these relations into eq. (7.3), it may be verified by the reader (see Problem 7.3) that ΔT_f is given by

$$\Delta T_f = \frac{W_f}{(G^2 + 4\pi^2 f^2 H^2)^{1/2}} \quad (7.4)$$

For good sensitivity, it is obviously desirable to have as large a value of ΔT_f as possible; inspection of eq. (7.4) shows that this implies small values for both H and G . This may be achieved by using thin absorbing elements of small area (to reduce H) which have minimal support (to reduce G). We may thus expect high sensitivity devices to be rather fragile. Looking now at the frequency characteristics, we may rewrite eq. (7.4) as

$$\Delta T_f = \frac{W_f}{G(1 + 4\pi^2 f^2 \tau_H^2)^{1/2}}$$

where τ_H , the *thermal time constant*, is given by

$$\tau_H = H/G \quad (7.5)$$

For good response at a frequency f we require $\tau_H \ll 1/2\pi f$.

Thus once H has been fixed (from size considerations) then G cannot be made too small, otherwise the response time may become unduly long. Typical values for τ_H found in practice usually range from 10^{-3} s upwards, although smaller values can be achieved. The limiting sensitivity of thermal detectors is governed by temperature fluctuations within the detector, which arise from random fluctuations in the energy flow rate out of the element. It may be shown (ref. 7.1) that the root mean square (r.m.s.) fluctuations in the power (ΔW_{TL}), flowing through a thermal link, which have frequencies between f and $f + \Delta f$, can be written

$$\Delta W_{TL} = (4kT^2 G)^{1/2} \Delta f \quad (7.6)$$

The smallest value of G obtainable is when energy exchange takes place by means of radiative exchange only. In this case, the minimum detectable power at room temperature for a 100 mm^2 area detector is about $5 \times 10^{-11} \text{ W}$ (see Problem 7.4). The best detectors available approach to within an order of magnitude of this figure at room temperature, and the performance of some can be improved further by cooling.

The receiving element is often in the form of a thin metal strip with a suitably absorbant surface coating such as 'gold black' (an evaporated film of gold which is uniformly absorbant at wavelengths from the UV well into the IR). Mounting the element in a vacuum enclosure

gives increased stability due to isolation from air movement, although window transmission can then be a problem if a wide wavelength range is desired. To overcome drift in the output caused by changes in the ambient temperature, we can take the difference in output of two identical detectors in close proximity, only one of which is exposed to the incident radiation.

When very large amounts of radiation are encountered (such as in the outputs from high power lasers), more massive detector elements are used; these are often in the form of stainless steel disks or cones.

Because of their relative unimportance in the field of optoelectronics we deal only briefly with a few of the better known types of thermal detector; for a more detailed discussion, the reader is referred to ref. 7.1.

7.2.1 Thermoelectric detectors

Thermoelectric detectors use the principle of the thermocouple (i.e. the Seebeck effect) whereby the heating of one junction between two dissimilar metals relative to the other causes a current to flow round the circuit which is proportional to the temperature difference between the junctions. In thermoelectric detectors one junction is used to sense the temperature rise of the receiving element whilst the other is maintained at ambient temperature, as shown in Fig. 7.3. A rather more sensitive detector may be made by connecting several thermocouples together in series; the device is then known as a *thermopile*. Efficient operation calls for materials with large electrical conductivities (to minimize Joule heating effects) and also small thermal conductivities (to minimize heat conduction losses). These two requirements are usually incompatible, and a compromise has to be reached. Although metals are most often used for the junction materials, certain heavily doped semiconductors can offer improved sensitivity, but these are generally less robust and give rise to constructional problems. The usefulness of thermoelectric detectors lies in their simplicity and their rugged construction.

7.2.2 The bolometer

In the bolometer the incident radiation heats a fine wire or metallic strip causing a change

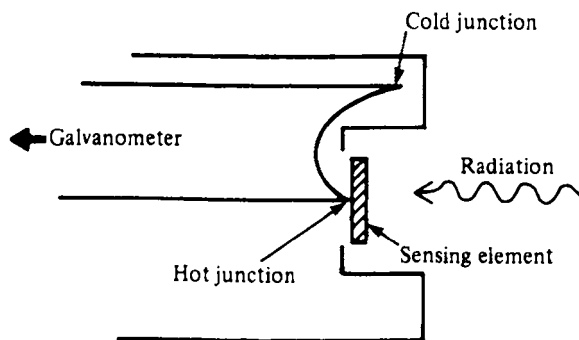


FIG. 7.3 Thermoelectric detector: the temperature change of the sensing element induced by the absorption of radiation is detected using a thermocouple, one junction of which is attached to the sensing element, and the other shielded from the radiation.

in its electrical resistance. This may be detected in several ways: for example, the element may be inserted into one arm of a Wheatstone bridge (Fig. 7.4), or in place of the photoconductor in the circuit of Fig. 7.17. Care must be taken to ensure that any currents flowing through the element are sufficiently small not to raise its temperature by a significant amount. The main parameter of interest in assessing the performance is the temperature coefficient of resistance α , which is given by

$$\alpha = \frac{1}{\rho} \frac{d\rho}{dT}$$

where ρ is the resistivity of the material and T the temperature. The resistivity of metals increases with increasing temperature, and hence for these α will be a positive quantity. Platinum and nickel are the most commonly used, and both have α values of about 0.005 K^{-1} . Greater sensitivity may be achieved by using semiconducting elements, which are sometimes called *thermistors*. These consist of oxides of manganese, cobalt or nickel and have α values of about -0.06 K^{-1} (for these materials, α is dependent on temperature). The negative sign arises because of the characteristic decrease in resistivity with increasing temperature of semiconductors above a certain temperature (see eq. 2.24 and the discussion following eq. 2.36).

Carbon resistance bolometers cooled to liquid helium temperature (4.2 K) have proved successful in far-IR astronomy where very sensitive detectors are required. Use can also be made of superconducting materials; in these the resistivity drops suddenly to zero below a particular temperature (the 'critical temperature'). In operation the temperature of the element is held just below this value so that the absorption of even a small amount of radiation will cause a very rapid increase in the resistivity. The devices exhibit exceptional sensitivity but

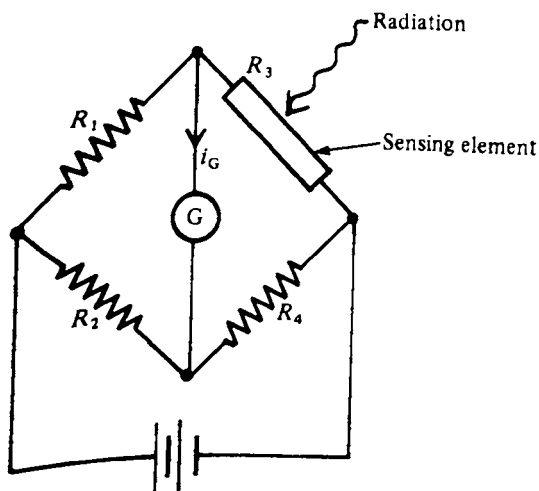


FIG. 7.4 Wheatstone bridge circuit incorporating a bolometer radiation sensing element. When the resistance values are such that $R_1/R_2 = R_3/R_4$ then the current i_G through the galvanometer is zero. If, however, the sensing element resistance changes slightly then a current will flow which is proportional to the resistance change.

require that the ambient temperature be very accurately controlled. The latest designs use high temperature superconducting materials such as yttrium barium copper oxide (YBCO).

7.2.3 Pneumatic detectors

The receiving element in a pneumatic detector is placed inside an airtight chamber. Radiation falling on the element causes the air temperature inside the chamber to rise and hence the air pressure to increase. This pressure increase may be detected in several ways. In one of the most sensitive detectors, the Golay cell, a wall of the chamber has a hole in it covered by a flexible membrane silvered on its outside surface. This acts as a mirror whose focal length depends on the pressure within the chamber (see Fig. 7.5). A beam of light originating from a source *S* passes through a grating, is then reflected from the flexible mirror to repass through the grating, and finally is directed onto a detector *D*. When no radiation is being absorbed within the chamber, the optics are arranged so that an image of the transmitting region of the grating is superimposed on a non-transmitting region and there is then no output from *D*. However, if the mirror changes its curvature slightly, light will be transmitted through the grating and recorded by *D*. The output from *D* is then proportional to the amount of radiation absorbed within the chamber. Golay cells are available which can detect radiation powers down to 10^{-11} W; they are, however, rather fragile and difficult to set up.

7.2.4 Pyroelectric detectors

Pyroelectric detectors are a more recent development, and while they do not have the same sensitivity as the Golay cell they can be made with very rapid response times and are very robust. The incident radiation is absorbed in a ferroelectric material which has molecules

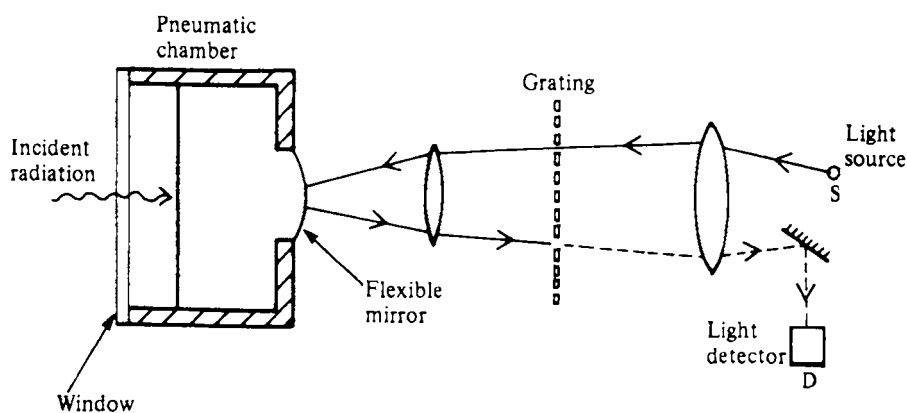


FIG. 7.5 Schematic diagram of a Golay cell detector. A beam of light originating from a source *S* passes through a grating. It is then reflected from a flexible mirror which forms part of the wall of a pneumatic chamber. The beam subsequently repasses through the grating and is directed onto a light detector *D*. Radiation absorbed within the chamber causes pressure fluctuations which in turn cause the curvature of the flexible mirror to change.

with a permanent electrical dipole moment. Below a critical temperature, the *Curie temperature* T_c , the dipoles are partially aligned along a particular crystallographic axis giving rise to a net electrical polarization of the crystal as a whole. When the material is heated, the increased thermal agitation of the dipoles decreases the net polarization, which eventually becomes zero above T_c , as shown in Fig. 7.6.

The most sensitive material in use is triglycine sulfate (TGS), but this has an inconveniently low Curie temperature of only 49°C and more commonly used materials are ceramic based, such as lead zirconate, which have Curie temperatures of several hundred degrees centigrade.

The detector consists of a thin slab of ferroelectric material cut such that the spontaneous polarization direction is normal to the large area faces. Transparent electrodes are evaporated onto these faces and connected together with a load resistor (which can have values as high as $10^{11} \Omega$) as shown in Fig. 7.7(a). A temperature change of the ferroelectric causes the spontaneous polarization to vary and hence also the amount of captive surface charge on the faces. Changes in the surface charge induce corresponding changes in the charge on the electrodes, thus causing a current to flow through the load resistor. This in turn results in a changing voltage signal appearing across the load resistor. Radiation of constant irradiance will not cause any change in the charge stored on the electrodes and will consequently not give rise to an output signal. The frequency response of the pyroelectric detector is considered in detail in Problem 7.6. At low frequencies, the output rises from zero to reach a plateau when $f > 1/2\pi\tau_H$, where τ_H is the thermal time constant given by eq. (7.5).

At higher frequencies, the electrode capacitance C acts as a signal shunt across the load resistor R_L and the output voltage falls to $1/\sqrt{2}$ of its maximum value at a cut-off frequency f_c given by

$$f_c = \frac{1}{2\pi R_L C} \quad (7.7)$$

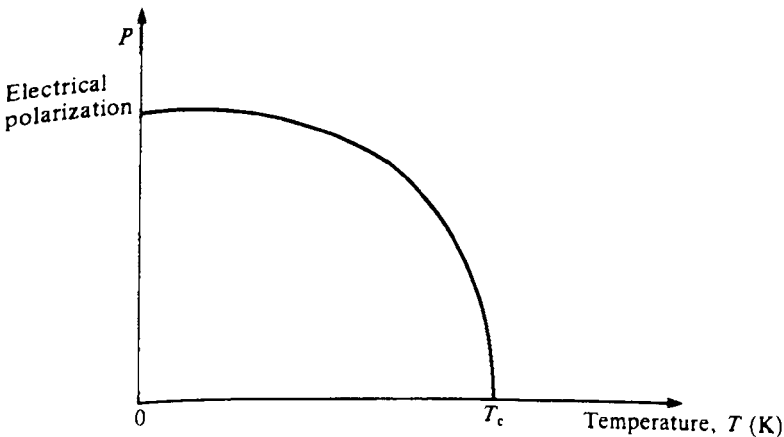


FIG. 7.6 Spontaneous electrical polarization versus temperature for a ferroelectric material (schematic diagram). The polarization falls to zero at the Curie temperature T_c .

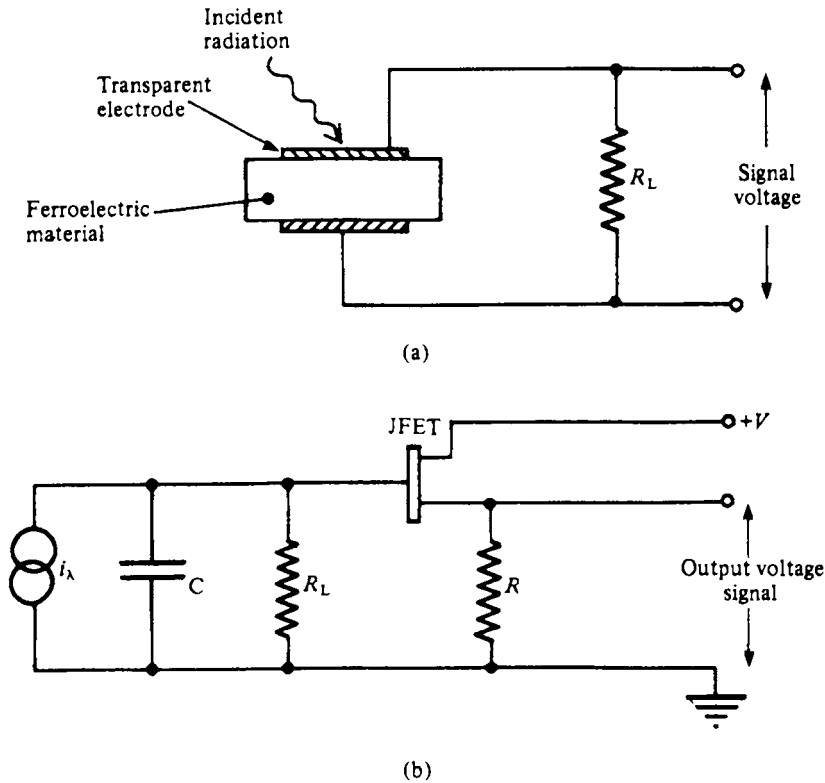


FIG. 7.7 (a) Pyroelectric detector. A slab of ferroelectric material is sandwiched between two electrodes (one being transparent). The electrodes are connected by a load resistor R_L . Radiation absorbed within the ferroelectric material causes it to change its polarization. The induced charge on the electrodes changes and current flows through R_L causing a voltage signal to appear across R_L . (b) Equivalent circuit and typical impedance matching circuitry for a pyroelectric detector. The varying amounts of charge stored on the electrodes are equivalent to a current generator feeding into the electrode capacitance C . The load resistor R_L is in parallel with C . Since R_L is usually very high (about $10^9 \Omega$ or more), an impedance matching circuit is often employed to reduce the signal source impedance. A typical circuit using a JFET is shown here; the output impedance in this case is then R ($\approx 1 \text{ k}\Omega$).

As the voltage output is proportional to R_L (again see Problem 7.6) there is a trade-off between sensitivity and frequency response. Typically, a detector with a frequency bandwidth of 1 Hz at an operating frequency of 100 Hz can detect radiation powers of about 10^{-8} W .

Because of the comparatively large values of the load resistor encountered in pyroelectric detectors, an impedance matching circuit is usually built into the detector. A source follower circuit using a JFET is commonly used as shown in Fig. 7.7(b).

Pyroelectric detectors can be made with response times in the nanosecond region and with a wavelength response extending out to $100 \mu\text{m}$. They have proved very useful as low cost, robust IR detectors in such uses as fire detection and intruder alarms.

7.3

Photon devices

7.3.1 Photoemissive devices

When radiation with a wavelength less than a critical value is incident upon a metal surface, electrons are found to be emitted; this is called the photoemissive or photoelectric effect (see section 2.6). When a photon of energy $h\nu$ enters the metal it may be absorbed and give up its energy to an electron. Provided the electron is then able to reach the surface and has enough energy to overcome the surface potential barrier (given by $e\phi$ where ϕ is the surface work function) it may escape and photoelectric emission takes place as illustrated in Fig. 7.8.

If the electron is initially at the Fermi level, its kinetic energy E on emission is given by

$$E = h\nu - e\phi \quad (7.8)$$

Since, however, the electron may be below the Fermi level initially and may also suffer inelastic collisions before emission, eq. (7.8) in fact represents the maximum energy available to the emitted electrons. No electrons at all will be emitted when $h\nu < e\phi$ (or, in terms of wavelength, $\lambda_0 > hc/e\phi$). If the probability of inelastic collisions of the excited electrons is high, then only a fraction of them may be able to escape. The ratio of the number of emitted electrons to the number of absorbed photons is called the *quantum yield* or *quantum efficiency*.

Pure metals, however, are rarely used as practical photocathodes since they have low quantum efficiencies ($\sim 0.1\%$), and high work functions. (Caesium has the lowest value for $e\phi$, namely 2.1 eV.) We may divide practical photoemissive surfaces into two groups: (a) the older *classical* types and (b) the newer *negative electron affinity* (NEA) types. The former consist of a thin evaporated layer containing compounds of alkali metals (usually including Cs) and one or more metallic elements from group V of the periodic table (e.g.

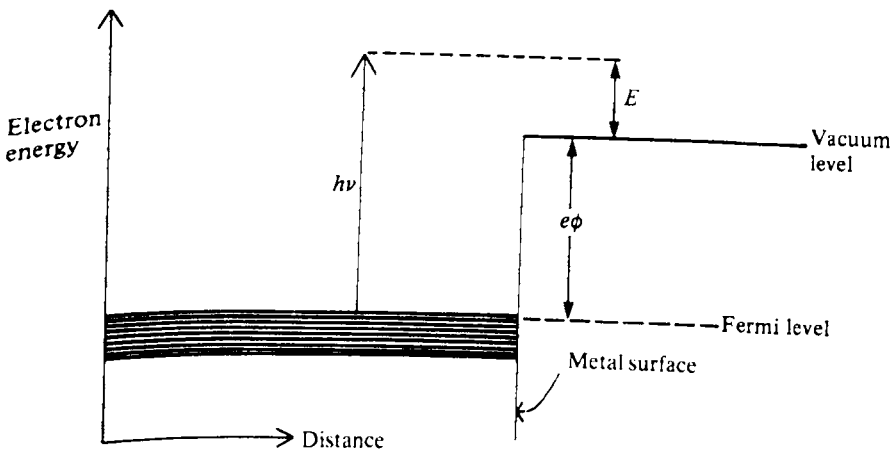


FIG. 7.8 Electron energy level diagram at a metal/vacuum interface illustrating the photoelectric effect. To escape from inside a metal, an electron must gain at least an energy $e\phi$, where ϕ is the work function expressed in electron volts.

Sb). They are often designated by an 'S' number. To some extent we may regard them as semiconductors, and hence most electrons must gain an energy of at least $E_g + \chi$ (where E_g is the energy gap and χ is the electron affinity) to escape from the surface (see Fig. 2.17b). For example, in the material NaKCsSb ('S20') $E_g = 1$ eV and $\chi \approx 0.4$ eV, and hence it should have a threshold at a photon energy of about 1.4 eV, which is indeed close to that observed.

It is possible, however, to reduce the effective value of χ , as far as the bulk electrons are concerned, provided that *band bending* takes place at the surface. This is found to occur when there are states within the energy gap at the semiconductor surface. If we suppose that a large number of these are close to the valence band, they will fill with holes and lead to a local depletion in the hole population. The resulting uncovering of negatively charged acceptor ions in the vicinity will lead to the formation of a depletion region at the surface very similar to that formed within the p material at a p-n junction. The potential drop across the depletion region leads to band bending as shown in Fig. 7.9. If we write the potential drop as V_s , then the effective electron affinity χ_{eff} , as far as the electrons in the bulk are concerned, is given by

$$\chi_{\text{eff}} = \chi - V_s$$

If $V_s > \chi$ then we have a negative electron affinity and the effective work function for bulk electrons is just E_g .

In practice, NEA photocathodes are formed by evaporating very thin layers of caesium or caesium oxide onto the semiconductor surface. The resulting band structure is more complex than shown in Fig. 7.9 but the essential features remain. Photocathodes using GaAs can be made which operate with quite high quantum efficiencies up to a wavelength corresponding to the energy gap of GaAs ($\approx 0.9 \mu\text{m}$). So far, it has not proved possible to

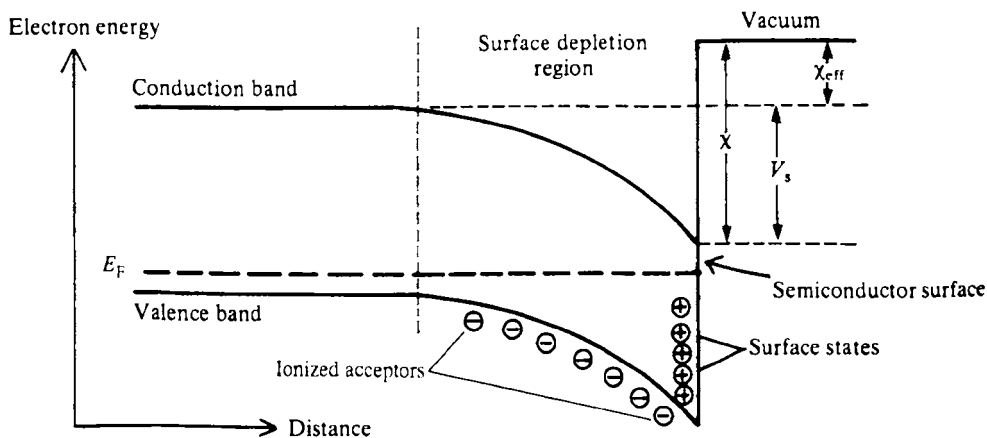


FIG. 7.9 The effective electron affinity of a semiconductor may be altered if band bending takes place at the surface. Here holes trapped in surface states cause a surface depletion region to be formed. The potential drop V_s across the depletion region reduces the effective electron affinity for bulk electrons from χ to $\chi - V_s$.

make NEA photocathodes which operate much above $1.1\ \mu\text{m}$ with any reasonable quantum efficiencies. There is no doubt, however, that the NEA photocathodes will prove to be superior to the 'classical' types in the near IR, although the lower cost of the classical types will ensure that they will continue to be used in the visible region. The quantum efficiencies of a number of the more common photocathode materials are shown as a function of wavelength in Fig. 7.10.

7.3.2 Vacuum photodiodes

In the vacuum photodiode, a photoemissive surface (the *photocathode*) is placed inside a vacuum tube with another electrode (the anode) placed nearby and biased positively with respect to it, as shown in Fig. 7.11. When the photocathode is illuminated, the emitted electrons will be collected by the anode and a current will flow in the external circuit. If the bias voltage is large enough (in practice a few hundred volts), all the emitted electrons will be collected and the current will be almost independent of bias voltage, but proportional to the light irradiance.

If monochromatic radiation with a (vacuum) wavelength of λ_0 and power P_λ is incident on a photocathode then the number of photons N_p incident per second is given by

$$N_p = \frac{P_\lambda}{hc/\lambda_0} = \frac{P_\lambda \lambda_0}{hc}$$

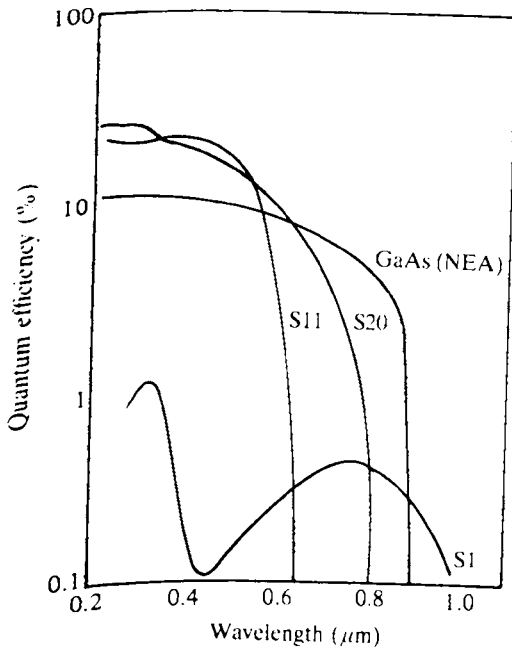


FIG. 7.10 Quantum efficiency versus wavelength for a number of the more common photocathode materials.

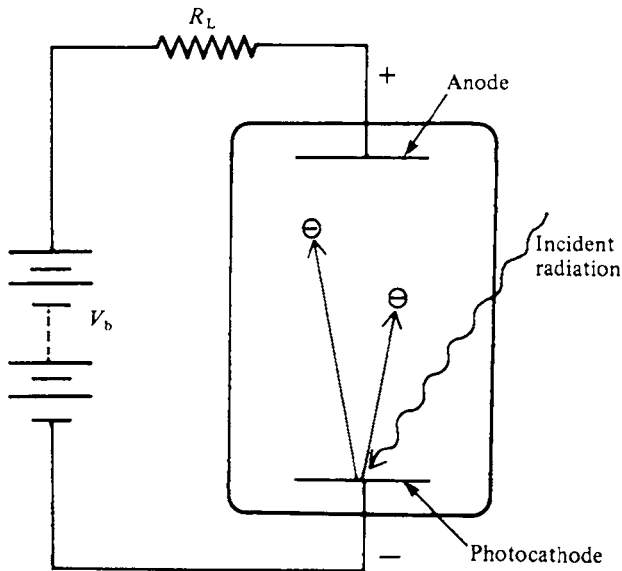


FIG. 7.11 Schematic diagram of a photoelectric cell. Electrons released from the photocathode surface by the incident radiation are attracted to the anode, thus causing a current to flow in the external circuit. A voltage will then appear across the load resistor R_L .

and hence if the quantum efficiency of the photocathode is η then the current flowing through the external circuit, i_λ , is given by

$$i_\lambda = \frac{\eta e P_\lambda \lambda_0}{hc} \quad (7.9)$$

Unless the incident power is relatively large the outputs from vacuum photodiodes are small (see Example 7.2) and thus often require external amplification (this relative insensitivity can be useful when examining high power laser pulses, however). A much more popular device is the *photomultiplier* which uses the same basic principles as the vacuum photodiode but which amplifies the resulting currents internally.

EXAMPLE 7.2 Vacuum photodiode output

Suppose that $1 \mu\text{W}$ of radiation with a wavelength of 500 nm is incident on a vacuum photodiode whose photocathode has a quantum efficiency of 0.5 . The current generated is then given by eq. (7.9) as

$$i_\lambda = \frac{0.5 \times 1.6 \times 10^{-19} \times 1 \times 10^{-6} \times 500 \times 10^{-9}}{6.6 \times 10^{-34} \times 3 \times 10^8} = 2 \times 10^{-7} \text{ A}$$

7.3.3 Photomultipliers

In the photomultiplier, the photoelectrons are accelerated towards a series of electrodes (called *dynodes*) which are maintained at successively higher potentials with respect to the cathode. On striking a dynode surface, each electron causes the emission of several secondary electrons which in turn are accelerated towards the next dynode and continue the multiplication process. Thus, if on average δ secondary electrons are emitted at each dynode surface for each incident electron and if there are N dynodes overall, then the total current amplification factor between the cathode and anode is given by

$$G = \delta^N \quad (7.10)$$

Considerable amplification is possible: if we take, for example, $\delta = 5$ and $N = 9$, we obtain a gain of 2×10^6 .

Four of the most common photomultiplier dynode configurations are illustrated in Fig. 7.12. Three of them (venetian blind, box and grid, and linear focused) are used in 'end-on' tubes. These have a semitransparent cathode evaporated onto the inside surface of one end of the tube envelope. The photoelectrons are emitted from the opposite side of the cathode layer to that of the incident radiation. Obviously, in this arrangement the thickness of the photocathode is very critical. If it is too thick, few photons will penetrate to the electron-emitting side, whilst if it is too thin few photons will be absorbed.

In the venetian blind type, electrons strike a set of obliquely placed dynode slats at each dynode stage; the electrons are attracted to the next set of slats by means of the interdynode potential applied between a thin wire grid placed in front of the slats. This arrangement is compact, relatively inexpensive to manufacture and is very suitable for large area cathodes. The box and grid type (Fig. 7.12b) is somewhat similar in performance. In both of these, very little attempt is made to focus the electrons, which is in contrast to the linear focused and circular cage focused types (Figs 7.12c and d), where some degree of electron focusing is obtained by careful shaping and positioning of the dynodes.

The focused types have somewhat higher electron collection efficiencies and a much better response to high signal modulation frequencies (we discuss frequency response later in this section). The circular cage focused type is very compact and usually used in conjunction with a side window geometry. In this, the photocathode material is deposited on a metal substrate within the glass envelope and the photoelectrons are emitted from the same side of the cathode as that struck by the incident radiation.

The dynode potentials are usually provided by means of the circuit shown in Fig. 7.13. Care must be taken to ensure that the voltage between the cathode and the first dynode is large enough to maintain proportionality between cathode current and cathode illumination. Usually a voltage value is recommended for a particular tube, and in some circumstances (e.g. when examining fast pulses) it may be preferable to use a Zener diode in place of the fixed resistor R_k in Fig. 7.13 to keep the voltage at this value.

The intermediate stages usually operate satisfactorily over quite a wide voltage range provided the voltage is distributed uniformly. To maintain this uniformity, the current flowing down the dynode chain must be considerably larger (say 100 times) than the anode current.

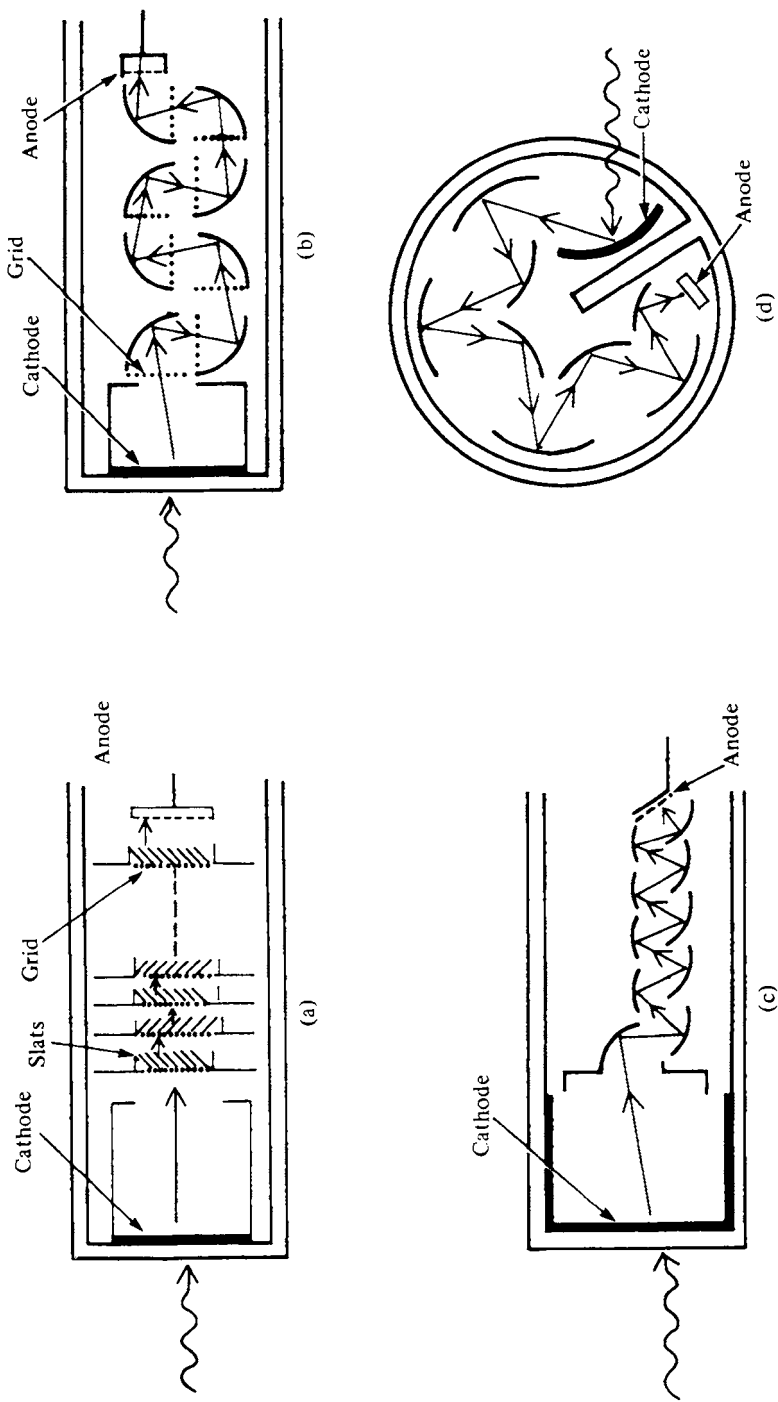


FIG. 7.12 Dynode structures of four common types of photomultiplier: (a) venetian blind, (b) box and grid, (c) linear focused and (d) circular cage focused. Typical trajectories of an electron through the systems are also shown.

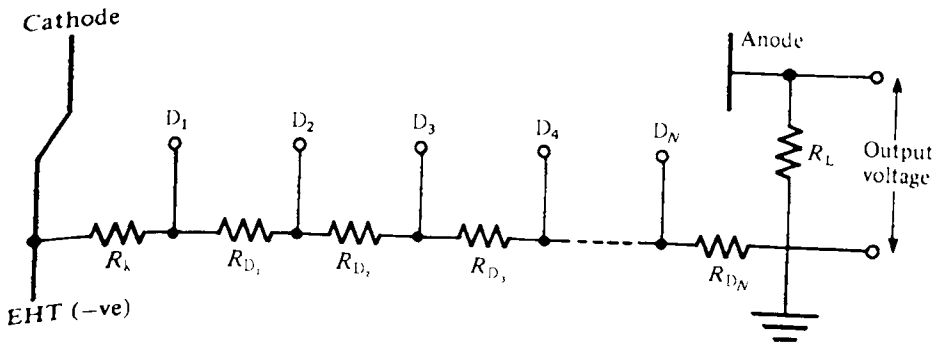


FIG. 7.13 Dynode biasing circuit using a linear resistor chain. The EHT voltage supply is applied across a resistor chain (R_k , R_{D1} , R_{D2} , etc.) which acts as a potential divider and maintains the dynodes (D_1 , D_2 , D_3 , etc.) at increasingly higher positive potentials relative to the cathode. When an amplified signal current pulse arrives at the anode, it flows through the anode load resistor R_L causing a voltage to appear across it.

If high anode currents are likely, then the last few stages may also be biased using Zener diodes.

The photomultiplier responds to light input by delivering charge to the anode. This charge may be allowed to flow through a resistor R_L or to charge a capacitor; the corresponding voltage signal then provides a measure of the input optical signal. If individual pulses need to be examined (as, for example, when using photon counting techniques, which we discuss in section 7.3.9), then it is important to ensure that the response time of the external circuitry is less than that of the pulse rise time. This usually implies a low value for the load resistor.

Traditionally photomultipliers have been relatively bulky devices with photocathode diameters of 25 mm or more. Recently much smaller devices have become available contained in small metal cans with photocathode diameters of 8 mm or so.

7.3.3.1 Speed of response

The electrons take a finite time (the *electron transit time*) to traverse the dynode chain from cathode to anode, although in fact this time will not be identical for all the electrons. There are two main reasons for the spread in transit times: first, the electrons have a spread in velocities when they are ejected from the cathode and subsequent dynodes; secondly, they traverse slightly different paths (in different electric fields) through the photomultiplier. It is this spread in transit times that limits the ability of the photomultiplier to respond faithfully to a fast optical pulse. Transit time spread may be reduced by using fewer dynodes and designing these so that electrostatic focusing gives the electrons very similar paths in nearly identical fields (as, for example, in the linear focused type, shown in Fig. 7.12c). Overall tube gain may be maintained when using fewer dynodes if the interdynode voltages are relatively high and if the dynode surfaces have high secondary electron emission coefficients. Dynode surface materials are now available which have much higher values of δ than that of the hitherto commonly used Be-Cu. Photomultiplier tubes with rise times of about 2 ns and transit times of about 30 ns are now readily available. In terms of the structures illustrated in Fig. 7.12, the order

of increasing speed of response is: box and grid, venetian blind, circular focused and linear focused.

7.3.3.2 Noise in photomultipliers

All electrical systems suffer from noise, that is a randomly varying output signal unrelated to any input signal. In a photomultiplier there are several sources of noise which we now discuss.

DARK CURRENT

Even when no radiation is falling onto the photocathode surface, thermionic emission gives rise to a 'dark current' which often constitutes the main source of noise in photoemissive devices. The thermionic emission current i_T for a cathode at temperature T of area A and work function ϕ is given by the Richardson–Dushman equation (ref. 7.2)

$$i_T = aAT^2 \exp\left(-\frac{e\phi}{kT}\right) \quad (7.11)$$

Here a is a constant which for pure metals has the value $1.2 \times 10^{-6} \text{ A m}^{-2} \text{ K}^{-2}$. Thermionic emission may be lowered by reducing the temperature and indeed this is often essential, especially for low work function photocathodes, although other sources of noise may dominate at low temperatures (e.g. electrons may be emitted from the photocathode by radioactive bombardment). If the dark current were absolutely constant, it would be a relatively simple matter to subtract it from the total output; however, it is itself subject to random fluctuations due to the statistical nature of thermionic emission. The r.m.s. variation in the current is given by the same equation as for shot noise which is discussed next.

SHOT NOISE

Shot noise is encountered whenever there is current flow and arises directly from the discrete nature of the electronic charge. Thus when a current flows past any point in a circuit the arrival rate of electrons will fluctuate slightly and gives rise to fluctuations in the current flow at that point. It may be shown (ref. 7.3) that the magnitude of the r.m.s. current fluctuations Δi_c with frequencies between f and $f + \Delta f$ is given by

$$\Delta i_c = (2ie\Delta f)^{1/2} \quad (7.12)$$

where i is the current flowing, which in the case of the photomultiplier is the sum of the dark current and the signal current leaving the photocathode.

The presence of dark-current shot noise sets a limit on the minimum detectable signal. We assume that if an optical signal results in an electrical output signal that is smaller than the noise signal, then it cannot be detected without further processing. We define the *responsivity* R_λ of the photomultiplier as i/W , where W is the optical power falling on the photocathode (see section 7.1). Therefore, the minimum detectable signal power in the presence of a thermionic dark current i_T is given by

$$W_{\min} = \frac{(2i_T e \Delta f)^{1/2}}{R_\lambda} \quad (7.13)$$

EXAMPLE 7.3 Minimum detectable signal for a photomultiplier

To calculate the minimum signal power from eq. (7.13) we need values for i_T , R_λ and Δf . The dark current may be estimated from eq. (7.11); thus if we assume a cathode area of 1000 mm^2 , a work function ϕ of 1.25 eV and a cathode temperature of 300 K (so that $kT/e = 0.025 \text{ eV}$) we have

$$i_T = 1.2 \times 10^6 \times 10^{-3} \times (300)^2 \exp(-1.25/0.025) \\ = 2 \times 10^{-14} \text{ A}$$

Next we calculate the responsivity R_λ using eq. (7.9). If we assume a quantum efficiency η of 0.25 at a wavelength of $0.5 \mu\text{m}$ we have

$$R_\lambda = \frac{\eta e \lambda_0}{hc} = \frac{0.25 \times 1.6 \times 10^{-19} \times 0.5 \times 10^{-6}}{6.6 \times 10^{-34} \times 3 \times 10^8} = 0.1 \text{ A W}^{-1}$$

Finally, if we take a bandwidth of 1 Hz , then from eq. (7.13) we obtain

$$W_{\min} = \frac{(2 \times 2 \times 10^{-14} \times 1.6 \times 10^{-19} \times 1)^{1/2}}{0.1} = 8 \times 10^{-16} \text{ W}$$

MULTIPLICATION NOISE

It is found that the current noise at the anode is always greater than that expected from shot noise alone (i.e. $G(2ie\Delta f)^{1/2}$). The reason for this is a statistical spread in the secondary electron emission coefficient about the mean value δ which causes the anode current noise to be increased by a factor $[\delta/(\delta - 1)]^{1/2}$ (ref. 7.4). This factor is only appreciable when δ is near to unity; for a typical value of $\delta = 4$ the noise current is increased by some 15%.

JOHNSON (OR NYQUIST) NOISE

Johnson noise arises because of the thermal agitation of charge carriers within a conductor; the random nature of this motion results in a fluctuating voltage appearing across the conductor. The r.m.s. value of this voltage, ΔV_J , having frequency components between f and $f + \Delta f$ across a resistance R at a temperature T , is given by (see ref. 7.5)

$$\Delta V_J = (4kTR\Delta f)^{1/2} \quad (7.14)$$

In a photomultiplier, such a noise voltage will appear across the anode load resistor (R_L in Fig. 7.13). In practice, Johnson noise is often smaller than the dark-current shot noise (see Example 7.4).

EXAMPLE 7.4 Noise in photomultipliers

Suppose we take a photomultiplier which has a load resistor of $10^3 \Omega$ at 300 K and a bandwidth Δf of 1 kHz . Then from eq. (7.14)

$$\Delta V_J = (4 \times 1.38 \times 10^{-23} \times 300 \times 1 \times 10^3)^{1/2} = 4.1 \times 10^{-9} \text{ V}$$

If the photomultiplier has a dark current of 10^{-14} A , then from eq. (7.12) there will be shot

noise current of

$$\Delta i_n = (2 \times 10^{-14} \times 1.6 \times 10^{-19} \times 10^3)^{1/2} = 1.8 \times 10^{-15} \text{ A}$$

If the photomultiplier gain is 10^7 and if we ignore any multiplication noise contribution, the shot noise voltage signal appearing across the load resistor is

$$\Delta V_j = 1.8 \times 10^{-15} \times 10^7 \times 10^3 = 1.8 \times 10^{-5} \text{ V}$$

7.3.4 Image intensifiers

As their name implies, image intensifiers are designed to boost very low intensity optical images to the point where they become useful. They can also act as wavelength 'down-converters', that is they can convert near-IR radiation into visible radiation. The primary image is formed on a photocathode surface and the resulting photoelectron current from each point on the image is then intensified either by increasing the energy of the individual electrons (in the so-called *first-generation* types) or by increasing the actual numbers of electrons (in *second-generation* types). The electrons subsequently fall onto a cathodoluminescent phosphor screen to produce the intensified image. A schematic diagram of a typical first-generation type is shown in Fig. 7.14.

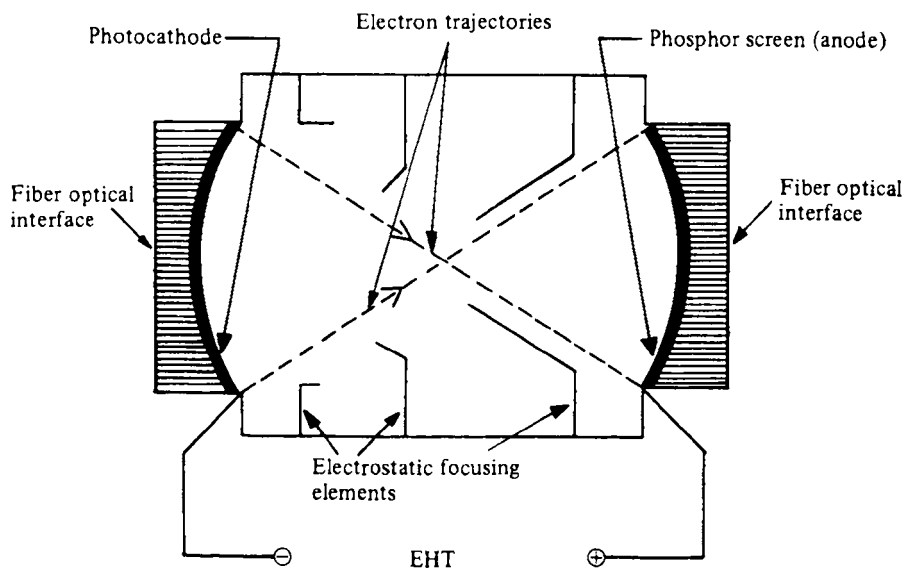


FIG. 7.14 Schematic diagram of a first-generation image intensifier. Electrons are released from the photocathode and accelerated towards the anode which is coated with a phosphor layer. Electrostatic focusing elements ensure that electrons released from a certain spot on the photocathode are all focused onto a corresponding spot on the phosphor screen. Both the photocathode and the phosphor screen are curved, and fiber optical coupling can be used to convert flat images into curved images and vice versa. On striking the phosphor screen, light is generated by cathodoluminescence.

The electrons may be focused onto the screen by either electrostatic or magnetic means; usually the former is used for reasons of simplicity. An accelerating potential is applied between the cathode and the phosphor screen, thus increasing the electron energy from a few electron volts to 10 keV or so. Luminance gains of up to 2000 may be achieved quite readily, while higher gains are possible by cascading two or more units with fiber optic coupling between them.

Second-generation devices make use of the so-called *microchannel plate* to achieve gain through electron multiplication. The microchannel plate consists of a slab of insulating material ($\approx 500\text{ }\mu\text{m}$ thick) with a high density of small diameter ($\approx 15\text{ }\mu\text{m}$) holes or *channels* in it. The inner surfaces of the channels are made slightly conducting and a potential ($\approx 1\text{ kV}$) is applied between opposite faces of the slab. Electrons entering one of the channels are accelerated down it and strike the walls soon after entering. Since the axis of the channel is slightly inclined to the electron trajectory, secondary electrons are generated by the impact and the process is repeated along the channel as illustrated in Fig. 7.15.

Focusing may be achieved most simply by placing the microchannel plate in close proximity to both the photocathode and the phosphor screen; this arrangement gives a very compact device. Alternatively, to achieve yet higher gain, the microchannel plate may be placed just in front of the screen in a first-generation image intensifier. In some instances an electronic output may be required rather than a visual one, in which case the luminescent screen may be replaced by an array of electron detectors such as silicon diodes.

A number of improvements to the basic second-generation scheme have resulted in what are termed *third-generation* devices. Use is made of high efficiency GaAs photocathodes which have peak spectral response in the $0.8\text{--}0.9\text{ }\mu\text{m}$ region where the night sky gives a photon flux some five to seven times that at a wavelength of $0.5\text{ }\mu\text{m}$. One of the problems

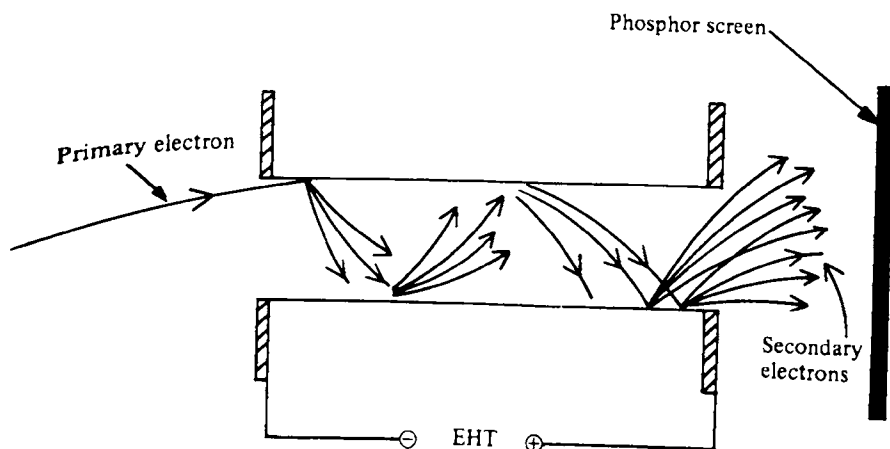


FIG. 7.15 Electron multiplication in a microchannel plate image intensifier. When the primary electrons enter the channel and strike the walls, secondary electrons are emitted that in turn generate further secondaries. The channel thus acts like a miniature photomultiplier tube. On emerging from the channel, the electrons generate light by striking a phosphor screen.

with the original microchannel plate is that positive ions generated within the microchannels bombard the photocathode eventually leading to its destruction. To counter this a thin membrane of aluminium oxide of thickness 3–5 nm is placed over the entrance to the microchannel plate which prevents any positive ions from reaching the photocathode. The membrane is sufficiently thin, however, to allow electrons with an energy greater than 200 eV to enter the channels. The newer plates also exhibit an improved spatial resolution by having smaller diameter channels (8 μm) and a smaller centre-to-centre channel spacing (10 μm). In addition the walls of the channels are coated with magnesium oxide, a secondary electron emission enhancer which further increases the gain.

7.3.5 Photoconductive detectors

As we showed in Chapter 2, an electron may be raised from the valence band to the conduction band in a semiconductor where the energy gap is E_g by the absorption of a photon of frequency ν provided that

$$h\nu \geq E_g$$

or in terms of wavelength

$$\lambda \leq hc/E_g \quad (7.15)$$

We define the *bandgap wavelength*, λ_g , to be the largest value of wavelength that can cause this transition, so that

$$\lambda_g = hc/E_g \quad (7.16)$$

As long as the electron remains in the conduction band, the conductivity of the semiconductor will be increased. This is the phenomenon of *photoconductivity*, which is the basic mechanism operative in photoconductive detectors. For convenience, we suppose the semiconductor material to be in the form of a slab of width W , length L and thickness D with electrodes on opposite ends, as shown in Fig. 7.16. An external potential across the electrodes is usually provided by the simple circuit of Fig. 7.17.

Any change in the conductivity of the detector results in an increased flow of current round the circuit which will increase the potential across the load resistor R_L . This may then be detected using a high impedance voltmeter. If we are only interested in the time-varying part of the incident radiation, then a blocking capacitor C may be inserted in the output line to remove any d.c. component. The optimum size for R_L in a particular situation is determined by the fractional change in the resistance of the photodetector when under maximum illumination. If this is small (say < 5%) then it may be shown (Problem 7.8) that the largest sensitivity is obtained when $R_L = R_D$, where R_D is the photodetector resistance. On the other hand, if it is relatively large, then linearity of output can only be maintained if the potential drop across the load resistor always remains small compared with the potential drop across the photoconductor. This requires that $R_L \ll R_D$. It is obviously advantageous, as far as the output voltage is concerned, that R_D should have a high value.

We suppose that the radiation falling normally onto the slab is monochromatic and of irradiance I_0 . Not all of this incident energy will be available to generate electrons within the

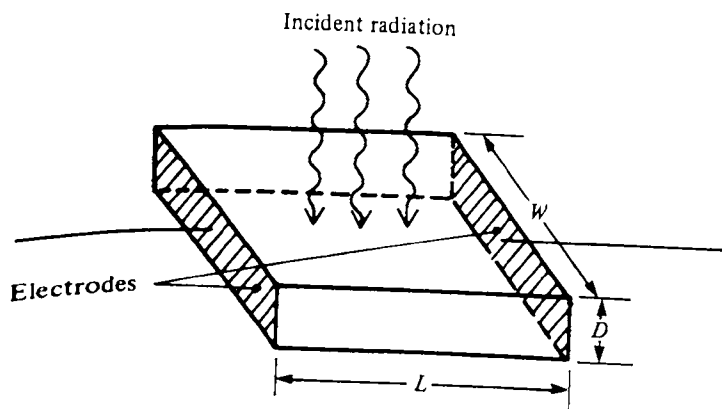


FIG. 7.16 Geometry of slab of photoconductive material. The slab of length L , width W and thickness D has electrodes on opposite faces; radiation falls onto the upper surface.

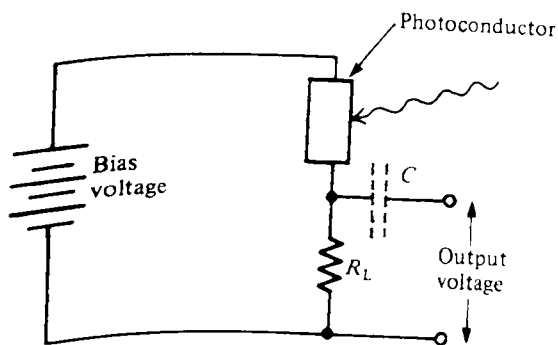


FIG. 7.17 Photoconductor bias circuit. The photoconductor is placed in a series circuit comprising a voltage source, a load resistor R_L and the photoconductor itself. Changes in the resistance of the photoconductor cause changes in the voltage appearing across R_L . If only the a.c. component of this voltage is required, then a blocking capacitor C may be placed as shown.

semiconductor: some will be reflected from the front surface, whilst some will pass through the semiconductor without being absorbed at all.

The reflection coefficient (assuming no antireflection layer is present) is given by r where

$$r = \left(\frac{n - 1}{n + 1} \right)^2$$

The irradiance just inside the surface of the slab is thus

$$I(0) = I_0(1 - r)$$

Now the irradiance at a point a distance x into the semiconductor, $I(x)$, can be written as

$$I(x) = I(0)\exp(-\alpha x)$$

where α is the *absorption coefficient* of the material. For wavelengths longer than the bandgap wavelength (λ_g), the absorption coefficient is comparatively small, whilst for wavelengths below λ_g , α increases rapidly with decreasing wavelength, and can attain values in the region of 10^6 m^{-1} or greater. Figure 7.18 shows the variation of α with wavelength for a number of important semiconductor materials.

Values of α of the order of 10^6 m^{-1} imply that most electron–hole pairs will be generated within a few micrometres of the semiconductor surface, although for wavelengths nearer the band edge this figure may be considerably larger. The fraction of the incident irradiance which is actually absorbed in the semiconductor can thus be written $(1 - r) \times \eta_{\text{abs}}$ where

$$\eta_{\text{abs}} = 1 - \exp(-\alpha D) \tag{7.17}$$

If we write $\eta = (1 - r) \times \eta_{\text{abs}}$ the total number of electron–hole pairs generated within the slab per second is $\eta I_0 WL/h\nu$. The average generation rate r_g of carriers per unit volume is then given by

$$r_g = \frac{\eta I_0 WL}{h\nu WLD}$$

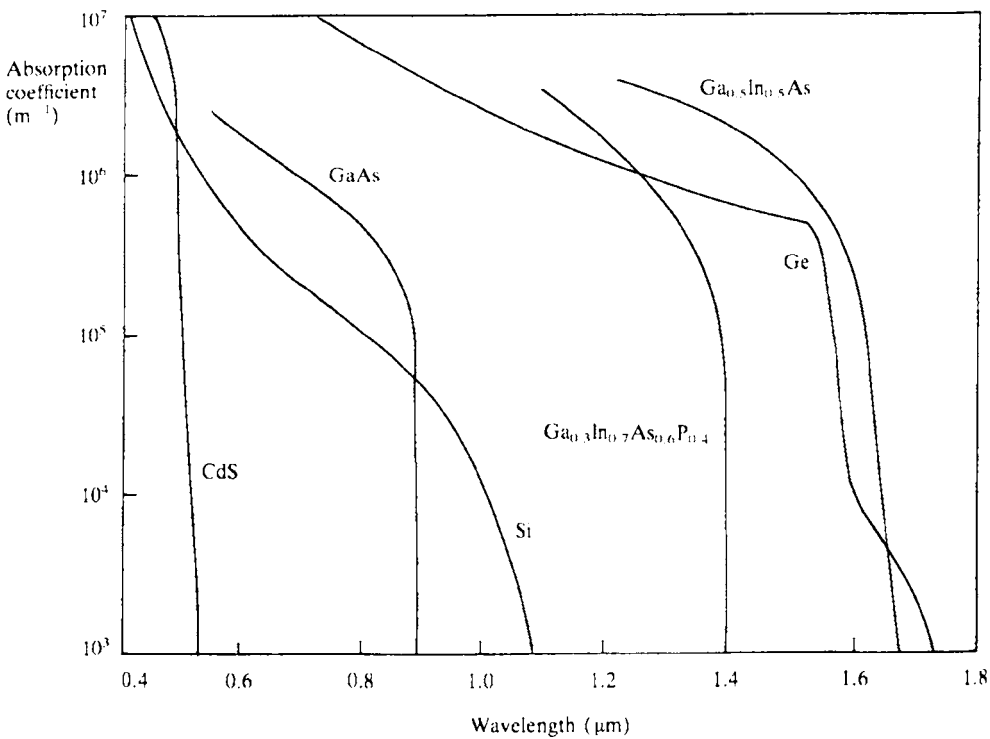


FIG. 7.18 Variation of the optical absorption coefficient α with wavelength for several semiconductor materials.

or

$$r_r = \frac{\eta I_0}{h\nu D} \quad (7.18)$$

As we saw in section 2.7 the recombination rate r_r of excess carriers depends on the densities of the excess carrier populations Δn , Δp (where for charge neutrality, $\Delta n = \Delta p$) and on the minority carrier lifetime τ_c via the equation

$$r_r = \Delta n / \tau_c = \Delta p / \tau_c \quad (7.19)$$

In equilibrium the recombination rate must equal the generation rate, and therefore

$$\Delta n = \Delta p = r_g \tau_c \quad (7.20)$$

We may write the conductivity σ of a semiconductor material as (see eq. 2.24)

$$\sigma = ne\mu_e + pe\mu_h$$

Hence under illumination the dark conductivity will increase by an amount $\Delta\sigma$ where

$$\begin{aligned} \Delta\sigma &= \Delta ne\mu_e + \Delta pe\mu_h \\ &= r_g \tau_c e (\mu_e + \mu_h) \end{aligned} \quad (7.21)$$

The application of a voltage V across the electrodes (Fig. 7.17) will result in a photoinduced current Δi given by

$$\Delta i = \frac{WD}{L} \Delta\sigma V$$

Using eq. (7.21) we then obtain

$$\Delta i = \frac{WD}{L} r_g \tau_c e (\mu_e + \mu_h) V \quad (7.22)$$

We may define an effective quantum efficiency parameter, known as the *photoconductive gain* G , as the ratio of the rate of flow of electrons per second from the device to the rate of generation of electron-hole pairs within the device. That is,

$$G = \frac{\Delta i}{e} \frac{1}{r_g WDL}$$

or, using eq. (7.22),

$$G = \frac{\tau_c (\mu_e + \mu_h) V}{L^2} \quad (7.23)$$

Unlike the quantum efficiency for, say, the photoelectric effect, G may be larger than unity. It may be increased by increasing V and decreasing L , although at high values of the electric field the current tends to saturate owing to space charge effects. High values of the gain will also be favoured by large values of τ_c , although this implies that the response time will

be correspondingly poor (see Problem 7.9). There is a good deal of evidence that in some materials (such as CdS) impurity energy levels exist within the energy gap into which carriers may fall but which do not cause recombination; the carrier is merely released by thermal excitation at some later time. Such levels are termed *traps* or *sensitization centres*. Whilst a carrier is held in a trap, a carrier of the opposite type must be present in the semiconductor to maintain charge neutrality – thus the presence of traps further increases the gain. This increase in gain, however, will once again be at the expense of the response time. Under fairly intense background illumination levels and at relatively elevated temperatures, the traps in most materials tend to be full; consequently they have little effect on the photosignal. Significant effects, however, may be observed at low temperatures and low background illumination levels.

Figure 7.19 shows an idealized wavelength response curve for a photoconductive detector assuming a constant quantum efficiency η for carrier generation when $\lambda < \lambda_g$. In practice, η is found to decrease at short wavelengths; this is a consequence of the increase in the absorption coefficient with decreasing wavelength (Fig. 7.18). Carriers are then being generated increasingly closer to the semiconductor surface where there is often a much higher probability of radiationless transitions taking place than in the bulk. Careful surface passivation techniques are needed to ensure good short wavelength response characteristics.

7.3.5.1 Noise in photoconductive detectors

The main source of noise in photoconductive detectors arises from fluctuations in the rates of generation and recombination of electron–hole pairs, and is termed *generation–recombination* noise. Both optical and thermal excitation processes contribute to generation noise. The relative importance of the thermal process is strongly dependent on the size of the bandgap, since the probability for thermal excitation of a carrier across the gap is approximately proportional to the factor $\exp(-E_g/2kT)$ (see the discussion following eq. 2.36). Thus for detectors capable of operating out to relatively long wavelengths, and which consequently have small energy gaps, thermal generation noise is likely to be large unless

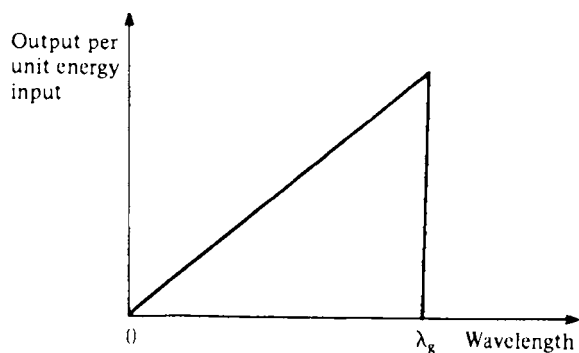


FIG. 7.19 Idealized wavelength response curve for a photoconductive detector. Beyond λ_g the output falls to zero since the photons have insufficient energy to excite carriers across the bandgap.

the temperature is reduced. As a rough rule of thumb, the operating temperature should be such that $T < E_g/25k$. Provided that the thermal generation noise has been reduced to negligible proportions the r.m.s. noise current fluctuations Δi_{gr} within the frequency range f to $f + df$ resulting from generation-recombination is given by (ref. 7.6)

$$\Delta i_{gr}(f) = \left(\frac{4ieG df}{1 + 4\pi^2 f^2 \tau_c^2} \right)^{1/2} \quad (7.24)$$

Here G is the photoconductive gain, τ_c the minority carrier lifetime and i the total current flowing.

The noise current given by eq. (7.24) is almost independent of frequency when $f \ll 1/(2\pi\tau_c)$; when $f > 1/(2\pi\tau_c)$ the noise current declines with increasing frequency. At sufficiently high frequencies, the dominant mechanism will be Johnson noise. The total noise present over a particular range of frequencies may be obtained by integration of eq. (7.24) (see Problem 7.10).

It is also found that at frequencies less than about 1 kHz a relatively little understood source of noise known as *flicker*, or $1/f$, noise becomes predominant. Flicker noise is present in most semiconductor devices and, although the cause has not been established with certainty, there appear to be some definite links with trap distributions that are metastable at the device operating temperature. Empirically it is found that the r.m.s. noise current due to flicker noise may be written as

$$\Delta i_{\text{flicker}}(f) = i \left(B \frac{df}{f} \right)^{1/2}$$

where B is a constant for a particular situation ($\approx 10^{-11}$). A complete noise spectrum for the photoconductive detector is shown schematically in Fig. 7.20.

7.3.5.2 Characteristics of particular photoconductive materials

Cadmium sulfide (CdS) and *cadmium selenide* (CdSe) are both used for low cost, visible radiation sensors, for example in light meters for cameras. These devices usually have high photoconductive gains (some 10^3 to 10^4) but poor response times (about 50 ms). The response time is in fact strongly dependent on the illumination level, being much reduced at high levels, a behaviour indicative of the presence of traps. A typical construction is shown in Fig. 7.21. A film of the material in polycrystalline form is deposited on an insulating substrate and the electrodes are formed by evaporating a suitable metal, such as gold, through a mask to give the comb-like pattern shown. This geometry, which results in a relatively large area of sensitive surface and a small interelectrode spacing, helps to give a high sensitivity to the device (see eq. 7.22).

Lead sulfide (PbS) is a well-known near-IR detector material with a useful wavelength response from 1 μm to 3.4 μm . By varying the growth conditions, the detector characteristics can be varied widely with the expected trade-off between gain and frequency response. A typical response time would be about 200 μs . In the wavelength region of 2 μm , it is one of the most sensitive photodetectors. The wavelength response may be extended to 4 μm

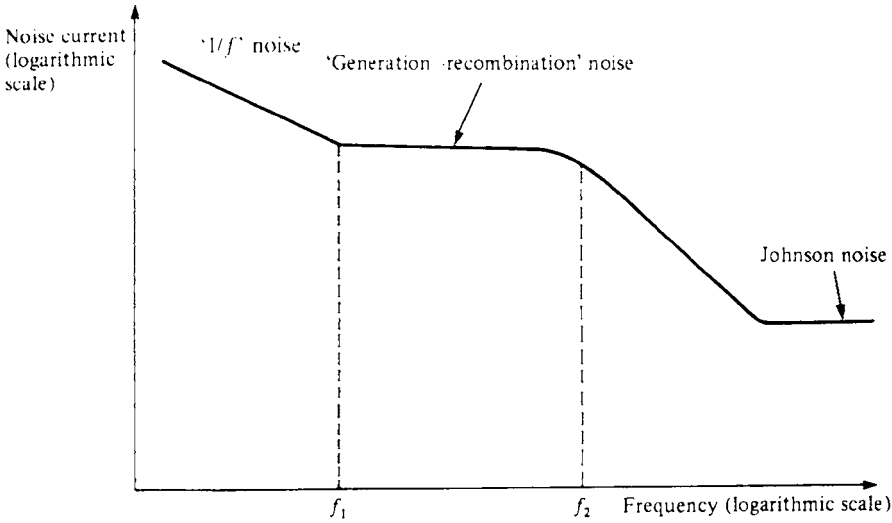


FIG. 7.20 Schematic diagram of the noise spectrum for a photoconductive detector. The frequency f_1 below which ' $1/f$ ' noise becomes predominant is about 1 kHz. Generation-recombination noise has a flat spectrum at frequencies below $1/(2\pi\tau_c)$ ($= f_2$) where τ_c is the minority carrier lifetime. Above f_2 generation-recombination noise falls with increasing frequency as $1/f$. At the highest frequencies (~ 1 MHz) Johnson noise associated with the circuit load resistor will eventually become predominant.

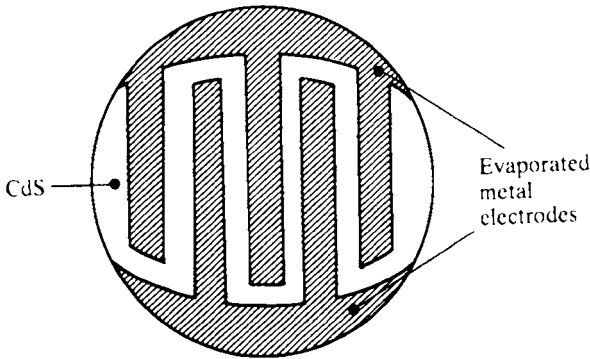


FIG. 7.21 Top view of typical electrode geometry of a CdS photoconductive cell.

by cooling to -30°C (unlike most other photoconductive materials, the energy gap of PbS decreases with decreasing temperature). However, cooling also has the effect of reducing both the overall gain and the frequency response. The internal dark resistance of these detectors is usually quite high ($\approx 1\text{ M}\Omega$). This is an advantage when using the bias circuit of Fig. 7.17, since then comparatively large values for the bias resistor R_L can be used, resulting in relatively high output voltage signals.

Indium gallium arsenide ($\text{In}_x\text{Ga}_{1-x}\text{As}$, $0 \leq x \leq 1$), the ternary alloy of InAs and GaAs, has proved a very useful detector material for the region $1\text{ }\mu\text{m}$ to $1.6\text{ }\mu\text{m}$. It may be readily grown

on InP substrates provided the two materials are lattice matched, which occurs when $x = 0.53$. The bandgap of $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ is 0.74 eV corresponding to a bandgap wavelength of $1.68 \mu\text{m}$ (see also Problem 7.13). Interdigitated electrodes similar to those illustrated for the CdS cell (Fig. 7.21) may be used. If these are spaced sufficiently closely together (a few micrometres) then quite small response times can be obtained (down to 5 ps or so), coupled with gains in the region of 50–100.

Indium antimonide (InSb) detectors are usually formed from single crystals and tend to have low impedances ($\approx 50 \Omega$). Consequently, their output voltages tend to be small. They have a wavelength response extending out to $7 \mu\text{m}$ and exhibit response times of around 50 ns. Operation at room temperature is possible, but a much improved noise performance results on cooling to liquid nitrogen temperatures (77 K), although the peak wavelength response then shifts to $5 \mu\text{m}$.

Mercury cadmium telluride ($\text{Hg}_x\text{Cd}_{1-x}\text{Te}$) may be thought of as an alloy composed of the semimetal HgTe and the semiconductor CdTe. Semimetals have overlapping valence and conduction bands and may be regarded as having a *negative* bandgap (for HgTe the bandgap is -0.3 eV). Consequently, depending on the composition of the alloy, a semiconductor can be formed with a bandgap varying between zero and 1.6 eV (the bandgap of pure CdTe). Detectors are available whose peak sensitivities lie in the range $5\text{--}14 \mu\text{m}$. This region is of particular importance, since it covers the peak emission wavelengths of bodies at or somewhat above ambient temperatures and also corresponds to a region of good atmospheric transmission (see Fig. 9.8). Usually cooling to 77 K or below is necessary for a satisfactory noise performance, but detectors operating at the lower end of this wavelength range are sometimes used in conjunction with a thermoelectric cooler. Photoconductive gains of up to 500 are possible, while at low temperatures and under low illumination the effects of traps can increase the gain yet further. As with indium antimonide, device resistances tend to be low.

DOPED SEMICONDUCTORS

Rather than using band-to-band transitions it is also possible to use transitions from impurity levels within the bandgap to the appropriate band edge. Typical among these are zinc- and boron-doped germanium detectors whose wavelength response can extend from $20 \mu\text{m}$ to $100 \mu\text{m}$, though cooling to 4 K (liquid helium temperatures) is essential to reduce background noise.

7.3.5.3 Vidicons and plumbicons

The vidicon is a generic name for a family of devices that relies on the phenomenon of photoconductivity to convert an optical image into an electrical signal. Figure 7.22 shows a typical structure. The optical image is formed on a thin target of semiconducting material (antimony trisulfide is commonly used) that has a transparent conducting layer (usually SnO_2) on the side facing the incident radiation. This conducting layer is connected to a potential of some +50 V above ground via a bias resistor. The other side of the semiconductor target is scanned with an electron beam in the same way as in a CRT. In operation, the target acts rather like a 'leaky' capacitor. When not illuminated its resistance will be

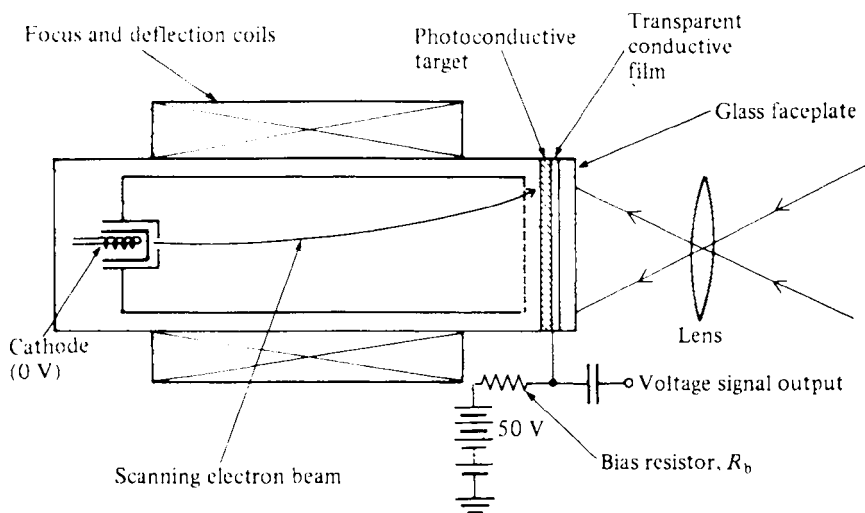


FIG. 7.22 Basic construction of a vidicon tube.

high, and charge will accumulate on opposite faces. The electron beam side will charge up to around cathode potential (0 V), whilst the other will charge up to around +50 V. Under illumination, however, the resistivity of the target material will be much reduced and the charge on the 'capacitor' will leak away (i.e. the capacitor will discharge itself) whenever the scanning beam is not incident on the area in question. When the beam does return to the 'discharged' area it will recharge the beam side and a corresponding amount of opposite charge must be supplied via the bias resistor and external bias supply circuit to the other side of the target. The amount of charge flowing will be dependent on how discharged the 'capacitor' has become, which, in turn, is directly related to the amount of light falling on the target. The output voltage signal is obtained by taking the voltage across the bias resistor.

One of the problems with the vidicon is its relatively high dark current, which gives rise to poor S/N ratios at low light irradiances. A device which exhibits very low dark currents is the plumbicon. This is essentially identical to the vidicon except for the nature of the photosensitive layer. In the plumbicon, this consists of a thin film p-i-n structure formed from lead oxide, PbO (Fig. 7.23a). The transparent SnO_2 layer acts as the n-type contact while the other surface has an excess of oxygen, which causes it to be p-type. The region in between (typically 15 μm thick) is effectively an intrinsic semiconductor. With no illumination, the potentials across the device cause it to be in reverse bias (and hence very little dark current will flow). Any illumination of the film generates electron-hole pairs, which will then flow to opposite sides of the structure and reduce the amount of stored charge. As in the vidicon, when the electron beam recharges the beam side a corresponding amount of charge is drawn through the bias resistor from the external supply. The plumbicon is widely used in colour TV studio cameras. The energy gap of PbO is about 2 eV so that the red sensitivity of the device is poor; this can be improved by adding a thin layer of PbS

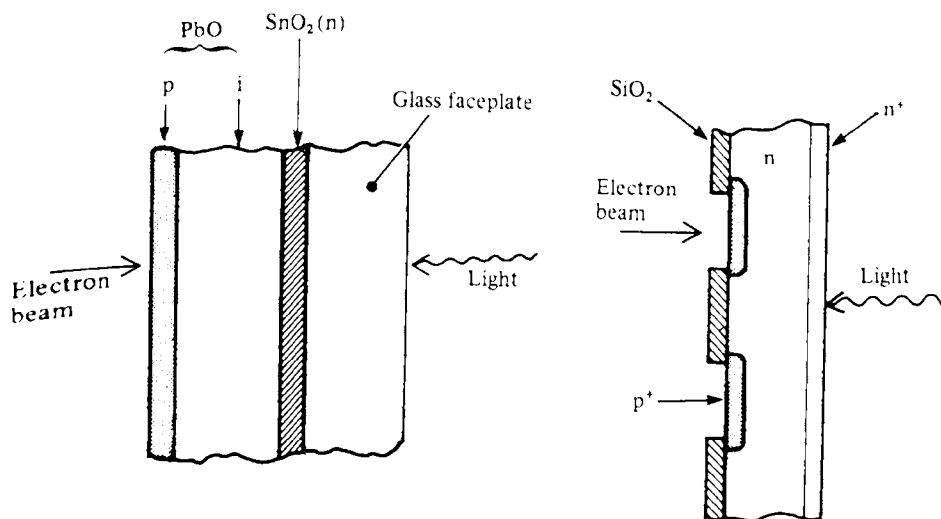


FIG. 7.23 Target structures for vidicon-type imaging tubes using (a) a lead-oxide-based p-i-n structure (the plumbicon) and (b) a discrete array of silicon p-n junction diodes.

($E_g = 0.4$ eV) to the electron beam side of the target. Any red light not absorbed by the PbO is then absorbed by the PbS.

A further development has been the replacement of the lead oxide target layer by an array of silicon diodes (Fig. 7.23b). Dark currents are very small, and the devices exhibit very uniform sensitivities over the wavelength range $0.45\text{--}0.85\text{ }\mu\text{m}$ with a good tolerance to high light levels.

7.3.5.4 Multiple quantum well detectors

The introduction of techniques for fabricating multiple quantum well (MQW) structures (see section 2.9) has led to the development of some very interesting detectors based on them which operate in the far-IR region. Within a quantum well an optically induced transition can take place between the sub-bands provided that the transition is 'vertical' on the E - k diagram (see section 4.6.1.1 and Fig. 4.13). That is, in the notation of section 2.9, a transition can take place between states of differing n_3 only if they have the same values of n_1 and n_2 (Fig. 7.24a). If we take an MQW structure and apply a field then the energy bands become 'tilted' and it is then possible for an electron in an excited state to tunnel through the potential barriers (Fig. 7.24b) and hence give rise to a current flow. A slightly different device can be made by using a well so narrow that only one sub-band is present within the well. Direct absorption can now take place between this state and the continuum states which lie above the top of the quantum well (Fig. 7.24c).

The sorts of transitions described here cover the wavelength range between about $5\text{ }\mu\text{m}$ and $20\text{ }\mu\text{m}$ (see Example 7.5). Although this range is already covered by the mercury cadmium telluride photoconductor there are considerable processing problems associated with making devices from this material.

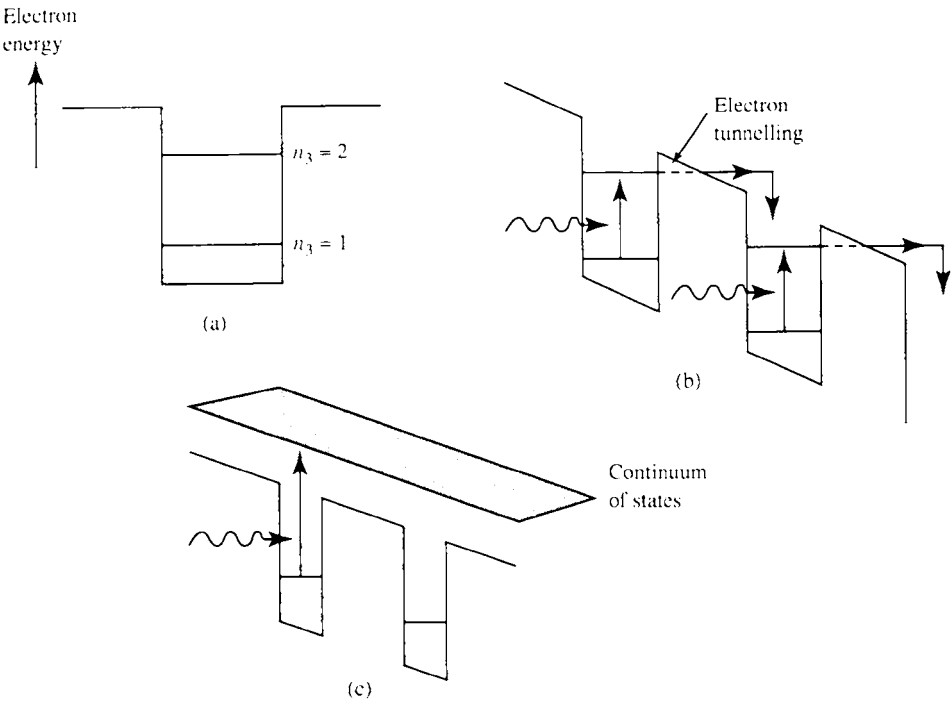


FIG. 7.24 (a) Electron states in a quantum well (no applied field); the dimensions are such that only two levels are present (quantum number $n_1 = 1$ and $n_2 = 2$). With an applied field we may observe photoconductivity due to (b) transitions between n_1 and n_2 followed by interwell tunnelling or (c) transitions to the electron continuum of states.

EXAMPLE 7.5 Detection wavelengths in quantum well detectors

According to eq. (2.64) the energy difference between the $n_1 = 1$ and $n_2 = 2$ states in a quantum well (n_1 and n_2 remaining constant) is given by

$$\Delta E = \frac{h^2}{8m_c^*} \left(\frac{2^2}{L_z^2} - \frac{1^2}{L_z^2} \right)$$

Here we have replaced m by m_c^* to allow for the fact that the well is within a semiconductor material and is not simply in a vacuum. Assuming a well 10 nm wide ($=L_z$) of GaAs (where $m_c^* = 0.068m$) we then obtain $\Delta E = 0.165$ eV. The corresponding optical wavelength is then given by $hc/\Delta E$, that is 7.5 μm .

7.3.6 Junction detectors

7.3.6.1 The p-n junction detector

When a p-n junction is formed in a semiconductor material, a region depleted of mobile

charge carriers is created which has a high internal electric field across it (see the discussion in section 2.8.1). If an electron-hole pair is generated by photon absorption within this region the electric field separates the electron and hole as shown in Fig. 7.25.

We may detect the charge separation in three distinct ways. First, if the device is left on open circuit (or with a very high resistance between the external contacts) an externally measurable potential will appear between the p and n regions; this is the *photovoltaic* mode of operation. Secondly, in the *photoamperic* mode a very low external resistance is connected between the external contacts and a photogenerated current flows through it. Finally the most usual way to operate the device is in the *photoconductive* mode where a reverse bias is applied across the junction and the resulting current flow through an external load resistor measured. The load resistor in this case need not be as small as in the photoamperic mode.

The junction will also respond to electron-hole pairs generated away from the depletion region provided they are able to diffuse to the edge of the depletion region before recombination takes place. From the discussion in Chapter 2, it is evident that only carriers generated within a minority carrier diffusion length or so of the edge of the depletion region are likely to be able to do this; nevertheless, this does increase the sensitive volume of the device.

In operation, we may represent the photodiode by a constant current generator (the current flow i_λ being generated by light absorption) with an ideal diode across it (to simulate the effect of the p-n junction), as shown in Fig. 7.26. The internal characteristics of the cell may be better modelled by the introduction of a shunt resistor (R_{sh}), a shunt capacitor (C_d) and a series resistor (R_s). If we assume a fraction η of the incident radiation is absorbed within the cell, then we may write

$$i_\lambda = \frac{\eta I_0 A e \lambda_0}{hc} \quad (7.25)$$

where I_0 is the light irradiance falling on a cell of area A .

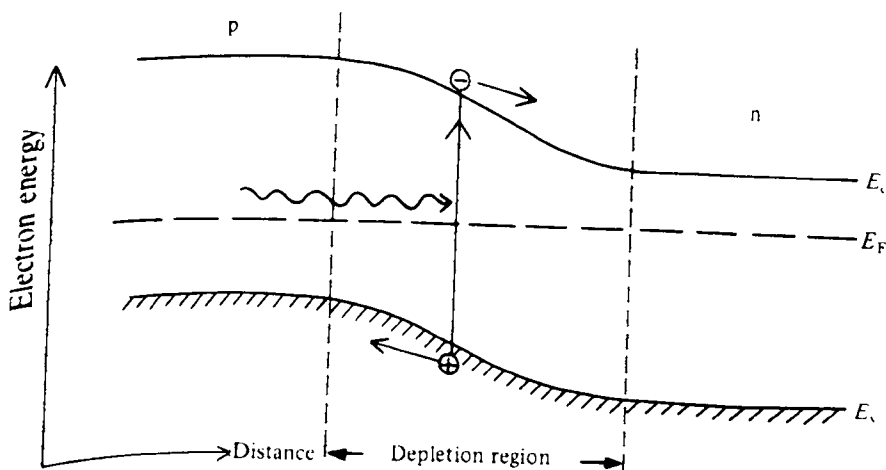


FIG. 7.25 Electron energy level diagram illustrating the generation and subsequent separation of an electron-hole pair by photon absorption within the depletion region of a p-n junction.

We now develop expressions for the sensitivities of the photodiode when operated in the photovoltaic and photoconductive modes. For simplicity, we assume that operation is at fairly low optical modulation frequencies, so that the effects of the shunt capacitance may be neglected (i.e. $i_c = 0$).

Considering the remaining current flows as shown in Fig. 7.26, we have

$$i_\lambda = i_d + i_{sh} + i_{ext} \quad (7.26)$$

also

$$V_{ext} = V_d - i_{ext}R_s \quad (7.27)$$

and

$$V_d = i_{sh}R_{sh} \quad (7.28)$$

Since the diode is assumed to be 'ideal' we may take its current-voltage behaviour to be given by eq. (2.51), that is

$$i_d = i_0 \left[\exp\left(\frac{eV_d}{kT}\right) - 1 \right]$$

where i_0 is the diode reverse bias leakage current.

In the photovoltaic mode, the external resistor (R_L) is made sufficiently large so that the current flowing through it, i_{ext} , is small compared with the photogenerated current, i_λ . In addition the value of the shunt resistor is usually sufficiently large that the shunt current i_{sh} is also much smaller than i_λ . Thus it is a reasonable approximation to assume that in eqs (7.26) and (7.27) we can put $i_{ext} = i_{sh} = 0$, and hence these equations become $i_\lambda = i_d$ and $V_{ext} = V_d$

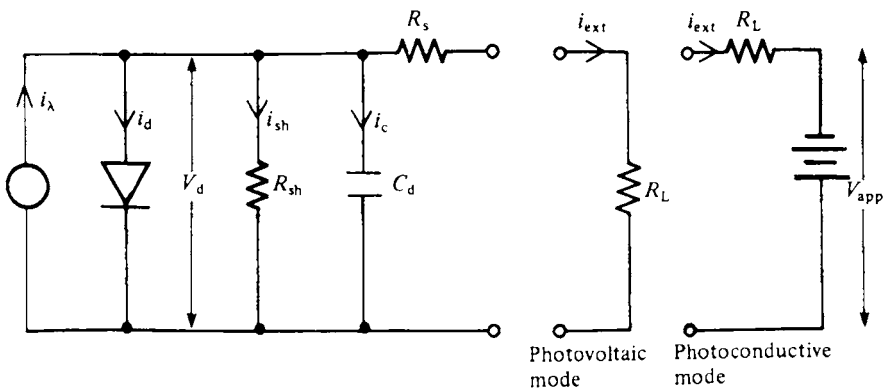


FIG. 7.26 Photodiode equivalent circuit. In operation, the photodiode can be represented by a photogenerated current source i_λ feeding into an ideal diode. The internal cell characteristics are better modelled by the introduction of a shunt resistor (R_{sh}), a shunt capacitor (C_d) and a series resistor (R_s). In the photovoltaic mode, an external high value resistor (R_L) is connected across the output and the voltage across this measured. In the photoconductive mode, an external bias (V_{app}) is applied in conjunction with a series load resistor (R_L). The current flowing through R_L is monitored by measuring the voltage across it.

respectively. Substituting for i_d in the former we obtain

$$i_\lambda = i_0 \left[\exp\left(\frac{eV_d}{kT}\right) - 1 \right]$$

Rearranging gives

$$\exp\left(\frac{eV_d}{kT}\right) = 1 + \frac{i_\lambda}{i_0}$$

Under normal operating conditions $i_\lambda \gg i_0$, so that we may write

$$\exp\left(\frac{eV_d}{kT}\right) = \frac{i_\lambda}{i_0}$$

It then follows that

$$V_d (= V_{\text{ext}}) = \frac{kT}{e} \ln\left(\frac{i_\lambda}{i_0}\right)$$

Substituting for i_λ from eq. (7.25) we have

$$V_{\text{ext}} = \frac{kT}{e} \ln\left(\frac{\eta I_0 e \lambda_0 A}{hc i_0}\right)$$

or

$$V_{\text{ext}} = \frac{kT}{e} \ln\left(\frac{\eta e \lambda_0 A}{hc i_0}\right) + \frac{kT}{e} \ln(I_0) \quad (7.29)$$

Hence the external voltage depends on the *logarithm* of the incident light irradiance.

In the photoconductive mode, a relatively large reverse bias (≈ 10 V or more) is usually applied across the diode (Fig. 7.26). Since the diode current saturates at i_0 for relatively small values of reverse bias (i.e. a few tenths of a volt), eq. (7.26) can be written as

$$i_\lambda = i_0 + i_{\text{sh}} + i_{\text{ext}}$$

As in the photovoltaic mode we may assume that both i_0 and i_{sh} are much less than i_λ and hence we may write $i_{\text{ext}} = i_\lambda$.

Substituting from eq. (7.25) we then have

$$i_{\text{ext}} = \frac{\eta I_0 A e \lambda_0}{hc} \quad (7.30)$$

With an external load resistor of R_L , the output voltage will then be

$$V_{\text{ext}} = \frac{\eta I_0 A e \lambda_0}{hc} R_L \quad (7.30a)$$

Hence, in the photoconductive mode the external current flowing is directly proportional to the incident light irradiance.

Finally, in the photoamperic mode, the situation is somewhat similar to that in the photoconductive mode: provided the external resistance is sufficiently small, almost all of the photogenerated current will flow through it and hence the output current and voltage will be the same as eqs (7.30) and (7.30a) respectively. However, if the external voltage ($=R_L \times i_{\text{ext}}$) becomes too large, the internal diode may become sufficiently forward biased to divert an appreciable amount of current from the external circuit; this will lead to a non-linear response.

The photoconductive mode thus offers the advantage of an inherently linear response over a relatively wide dynamic range. It also offers a more rapid response than the photovoltaic mode. The main drawback is the presence of a dark current ($i_0 + i_{\text{sh}}$) which, as in the photomultiplier, gives rise to shot noise (section 7.3.3.2) and limits the ultimate sensitivity of the device. All modes of operation are subject to generation noise (section 7.3.5.1), but recombination noise is absent since the charge carriers are separated in the depletion region before they can recombine.

A typical structure for a p-n diode junction detector based on silicon is shown in Fig. 7.27. Contact to the semiconductor material is made via a metal-n⁺ (or -p⁺) junction; this is found to be the most convenient way of providing an ohmic contact (see section 2.8.6). If we assume an abrupt junction together with the external bias, V , being much larger than the internal junction potential, V_0 , and also that $N_a \gg N_d$, then the depletion layer widths can be written as (see eqs 2.60, 2.59 and 2.55)

$$x_n = \left(\frac{2\epsilon_0\epsilon_r V}{eN_d} \right)^{1/2} \quad \text{and} \quad x_p = \left(\frac{2\epsilon_0\epsilon_r V N_d}{eN_a^2} \right)^{1/2} \quad (7.31)$$

Since we have a p⁺-n structure it follows that $x_n \gg x_p$; Fig. 7.28 illustrates the resulting electric field variation within the depletion regions. For efficient detection the electron-hole pairs should be generated within the depletion region. At short wavelengths, where the absorption

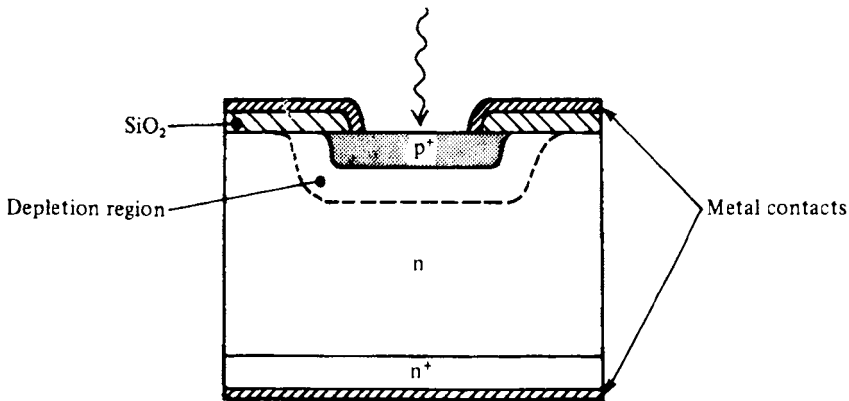


FIG. 7.27 Typical silicon photodiode structure for photoconductive operation. A junction is formed between heavily doped p-type material (p⁺) and fairly lightly doped n-type material so that the depletion region extends well into the n material. The p⁺ layer is made fairly thin. Metallic contacts can be made directly to the p⁺ material but to obtain an ohmic contact to the n material an intermediate n⁺ layer must be formed.

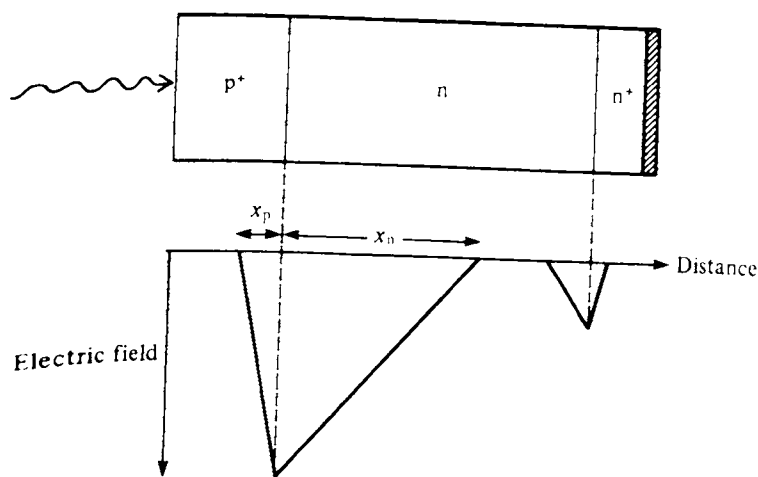


FIG. 7.28 Electric field distribution within the $p^+ - n$ junction diode shown in Fig. 7.27, assuming an abrupt diode structure.

coefficient is relatively high (Fig. 7.18), they will be generated close to the surface. Consequently, to achieve a good short wavelength response the p^+ region should be made as thin as possible. Conversely, at the upper wavelength range of the detector the absorption coefficient will be relatively small and a wide depletion region is required for high detection efficiency. This implies that large values of the reverse bias voltage are needed (see Example 7.6) which may approach or exceed the diode breakdown voltage. Detection efficiency may also be improved by providing an antireflection coating to the front surface of the detectors consisting of a $\lambda/4$ thick coating of SiO_2 .

EXAMPLE 7.6 Depletion region thickness in Si $p^+ - n$ photodiodes

If we take a silicon diode with a moderately doped n region, that is $N_d = 5 \times 10^{21} \text{ m}^{-3}$, and an applied voltage of 100 V, eq. (7.31) gives $x_n = 5.1 \times 10^{-6} \text{ m}$, that is $5.1 \mu\text{m}$.

Inspection of Fig. 7.18 shows that at $0.8 \mu\text{m}$ the absorption coefficient of silicon is about 10^5 m^{-1} , so that, ignoring Fresnel reflection, the fraction of the incident radiation absorbed by a $5.1 \mu\text{m}$ thick layer is $1 - \exp(-5.1 \times 10^{-6} \times 10^5)$, that is 0.4. This is obviously insufficient if a high efficiency photodiode is required at this wavelength. In fact to absorb 80% of the radiation would require the depletion layer thickness to be at least $20 \mu\text{m}$ wide. To achieve this with the doping levels considered here would require an applied voltage of approximately $100 \times (20/5)^2 \text{ V}$, or 1600 V! In fact electrical breakdown would occur long before a voltage as high as this could be applied.

7.3.6.2 The $p-i-n$ photodiode

A structure that results in a good long wavelength response with only relatively modest bias levels is the so-called $p-i-n$ (or PIN) structure, illustrated in Fig. 7.29. Here the intrinsic (i)

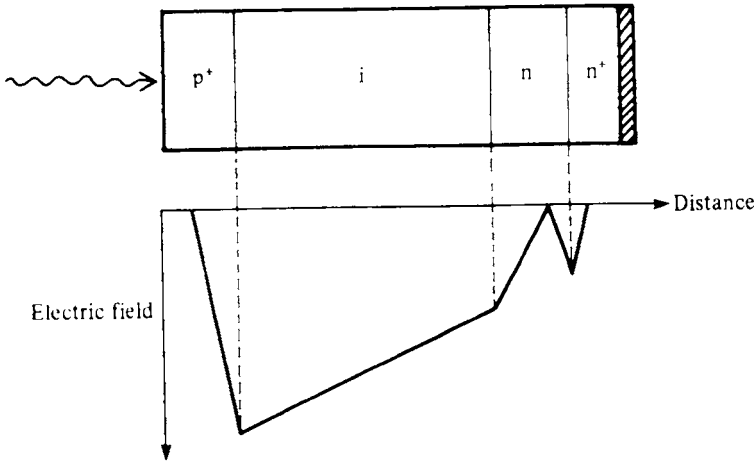


FIG. 7.29 Electric field distribution within a p-n structure.

region has a high resistivity (i.e. low values of N_a and N_d) so that only a few volts of reverse bias are needed to cause the depletion region to extend all the way through to the n region and thus provide a large sensitive volume. In practice, the bias is maintained at a considerably higher voltage than the minimum value and the intrinsic region then remains fully depleted of carriers even at high light levels. The depletion region width in a p-i-n structure is then practically independent of applied voltage and thus much better delineated than in a p-n structure where the depletion region width will vary appreciably with applied voltage. For this reason most simple photodiode structures are of the p-i-n rather than the p-n type.

For efficient detection of photons we require that as many as possible are absorbed within the intrinsic region (assuming a p-i-n structure). If the thicknesses of the p and i regions are w_p and w_i respectively, and assuming a surface reflectance of R , then the fraction F_i of the incident power that is absorbed within the i layer is given by

$$F_i = (1 - R) \{ \exp(-\alpha w_p) - \exp[-\alpha(w_p + w_i)] \} \quad (7.32)$$

Assuming that $w_p \ll \alpha^{-1}$, so that absorption within the surface p⁺ layer may be neglected, we may ensure that most of the incident radiation is absorbed in the i layer by requiring that $w_i \ll \alpha^{-1}$. To be more specific, if $w_i = 2\alpha^{-1}$ then some 86% of the radiation entering the device will be absorbed (there will also be reflection losses at the surface but these can be reduced by the incorporation of an antireflection coating). Silicon p-i-n photodiodes can achieve quantum efficiencies of up to 80% in the wavelength range 0.8–0.9 μm . A typical spectral response of a p-i-n silicon photodiode is shown in Fig. 7.30.

The problem of low detector efficiencies at wavelengths close to the bandgap wavelength can also be addressed by using detectors that are illuminated from the side (i.e. parallel to the junction) although these are not commonly available.

7.3.6.3 Photodiode materials

For the detection of radiation in the visible and near-IR region one of the most popular

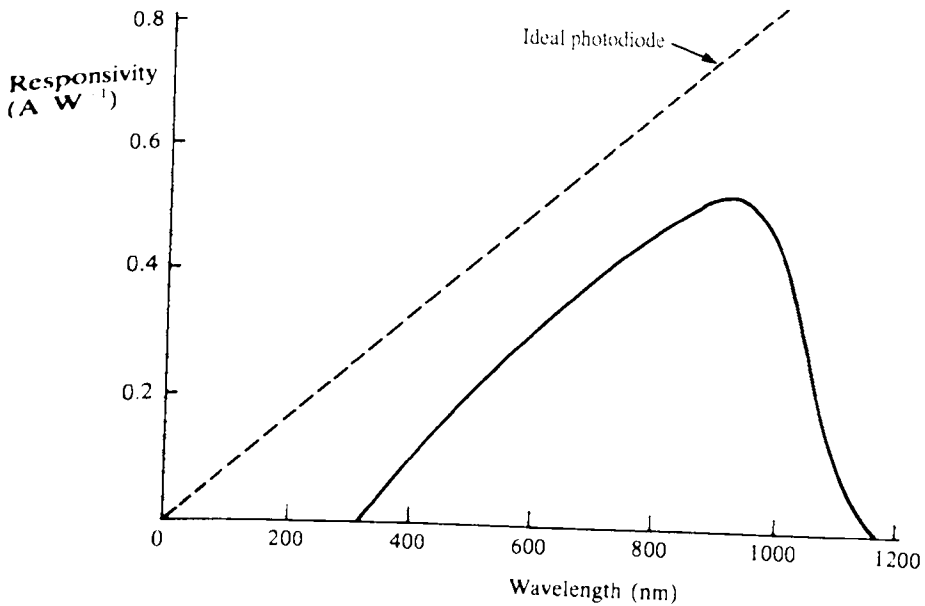


FIG. 7.30 Typical current responsivity of a silicon photodiode. Also shown is the responsivity of an ideal photodiode with unit quantum efficiency.

photodiode materials is silicon. This has an energy gap of 1.14 eV and, as mentioned above, can achieve quantum efficiencies of up to 80% between 0.8 μm and 0.9 μm . An examination of Fig. 7.18 shows that at these wavelengths the absorption coefficient is about 10^5 m^{-1} . Thus to obtain efficiencies as high as 80% we require an intrinsic layer thickness of about 20 μm and also a surface p^+ layer thickness that is substantially less than α^{-1} , or 10 μm .

As we shall see in Chapter 9, the development of optical fiber communication systems has led to the demand for detectors operating at wavelengths of 1.3 μm and 1.55 μm which have high sensitivities and exceptionally wide bandwidths. Photodiodes based on germanium are available (the bandgap wavelength for Ge is 1.88 μm) but they suffer from rather low sensitivities and large reverse bias leakage currents, this latter being a general problem in small bandgap semiconductors. Much more successful have been materials that can be grown on InP substrates such as the ternary compound $\text{In}_x\text{Ga}_{1-x}\text{As}$ and the quaternary compound $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$. The materials must be lattice matched to InP which, in the case of $\text{In}_x\text{Ga}_{1-x}\text{As}$, takes place when $x = 0.53$. The bandgap is then 0.74 eV which corresponds to a bandgap wavelength of 1.68 μm . Unfortunately when homojunctions are made from such narrow bandgap materials they display relatively low breakdown voltages and large reverse bias leakage currents. In fact by taking eq. (2.51a) and using the relationships $p_n \times n_n = n_i^2 = n_p \times p_p$ together with eq. (2.36), it can be shown that $J_0 \propto \exp(-E_g/kT)$. Consequently instead of homojunctions, it is usual to form heterojunctions with a wider bandgap material. An example of such a structure is shown in Fig. 7.31, where an i layer of $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ is sandwiched between a p^+ layer of InP and an n layer of InP. Because the radiation has to pass through a layer of InP which has a bandgap wavelength of 0.92 μm ,

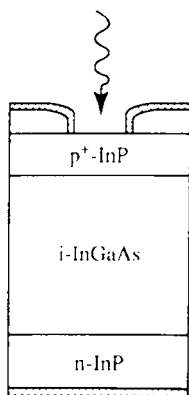


FIG. 7.31 Basic structure of a p-i-n heterojunction InGaAs detector.

then no radiation below this wavelength will be detected. A further advantage of this structure is that since no carriers of interest will be generated in either the surface p^+ layer or the lower n layer there will be no diffusion of carriers to the junction from outside the depletion region, which, as we shall see in the next section, can worsen the response time of the detector.

From Fig. 7.18 we can see that the absorption coefficient for $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ close to the bandgap wavelength is higher than in the case of silicon. This is essentially because $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ is a direct bandgap semiconductor whereas silicon is an indirect bandgap semiconductor. As a consequence the intrinsic layer thickness can be much smaller than in the case of silicon. At $1.5\ \mu\text{m}$, for example, the absorption coefficient of $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ is about $10^6\ \text{m}^{-1}$ and thus to achieve the condition $w_i = 2\alpha^{-1}$ we require an intrinsic layer thickness of the order of only $2\ \mu\text{m}$.

7.3.6.4 Response time of photodiodes

There are two main factors which limit the speed of response of photodiodes: these arise from the finite transit time of carriers across the depletion region and from the RC time constant associated with the electrical parameters of the diode and its external circuitry.

DRIFT TIME OF CARRIERS THROUGH THE DEPLETION REGION

In very high electric fields the drift velocities of carriers in semiconductors tend to saturate. Thus provided the field within the depletion region exceeds the saturation value for most of its length, then we may assume that the carriers move with the constant saturation velocity v_{sat} . Consequently they will take a time τ_{drift} to traverse the depletion region of width w , where

$$\tau_{\text{drift}} = w/v_{\text{sat}} \quad (7.33)$$

During the whole of the time that the carriers are moving across the depletion region a

current must be flowing through the external circuit. Thus if a very short duration pulse of radiation is absorbed within the depletion region then an output signal extending over a time τ_{drift} will be observed. Typical values for the saturation velocities are of the order of 10^5 m/s and these are attained in fields of above about 10^6 V m⁻¹.

JUNCTION CAPACITANCE EFFECTS

In section 2.8.4 it was shown that the capacitance of an abrupt junction can be written

$$C_j = A\epsilon_0\epsilon_r/w \quad (7.34)$$

where w is the total depletion region width. For a p-i-n photodiode we may assume a similar expression where w is now the width of the i region. Although junctions are rarely abrupt in practice, it still remains true that junction capacitance decreases with increasing reverse bias. For example, in a linearly graded junction we have $C_j \propto V^{-1/3}$.

We see from Fig. 7.26 that at high frequencies the diode capacitance acts as a shunt across the output resistance network and reduces the output. Usually $R_{\text{sh}} \gg R_L$ and $R_s \gg R_L$; hence as far as the diode current i_λ is concerned, the diode capacitance C_j is in parallel with the load resistor R_L . It is easy to show (Problem 7.14) that if the current i_λ is amplitude modulated at frequency f then the output voltage of the device is given by

$$V_o = \frac{i_\lambda R_L}{(1 + 4\pi^2 f^2 C_j^2 R_L^2)^{1/2}}$$

The electrical bandwidth Δf_{el} is defined as the frequency range over which the output is above $1/\sqrt{2}$ of its maximum value (see Appendix 5). Thus in the present instance we have $4\pi^2 \Delta f_{\text{el}}^2 C_j^2 R_L^2 = 1$, that is

$$\Delta f_{\text{el}} = \frac{1}{2\pi R_L C_j} \quad (7.35)$$

The bandwidth may obviously be improved by reducing C_j . Inspection of eq. (7.34) shows that this may be achieved by decreasing the diode area and increasing the thickness of the i region. Unfortunately both these courses of action have associated difficulties. There is obviously a limit on how small the diode area can be made without encountering problems associated with focusing an incident beam onto a small area. In addition increasing the width of the i layer will increase the value of τ_{drift} which will worsen the frequency response as far as carrier transit time is concerned. There will thus be an optimum value for the thickness of the i region which will result in maximum overall bandwidth (or, equivalently, a minimum in the overall response time).

Now the response time associated with the detector RC network can be written $\tau_{RC} (= R_L C_j)$ and the total response time, τ , including both transit time effects and junction capacitance effects can then be written²

$$\tau^2 = \tau_{\text{drift}}^2 + \tau_{RC}^2 \quad (7.36)$$

Now $\tau_{\text{drift}} \propto w$ and $\tau_{RC} \propto 1/w$ and it is easy to show (see Problem 7.16) that τ is minimized when $\tau_{\text{drift}} = \tau_{RC}$. This is the normal design criterion for ensuring the fastest diode response.

EXAMPLE 7.7 Response time of an InGaAs p-i-n photodiode

We consider a p-i-n photodiode based on InGaAs (i.e. as in Fig. 7.31) where the i region has a thickness of $2\text{ }\mu\text{m}$ and whose area is $100\text{ }\mu\text{m} \times 100\text{ }\mu\text{m}$. The load resistor used is $50\text{ }\Omega$. We take the saturation velocity of electrons in InGaAs to be 10^5 m s^{-1} and the relative permittivity of InGaAs to be 12.

The transit time of electrons through the depletion region (τ_{drift}) is given by $2 \times 10^{-6}/10^5\text{ s}$, or $2 \times 10^{-11}\text{ s}$.

From eq. (7.34), the device capacitance is given by

$$C_j = \frac{(100 \times 10^{-6})^2 \times 8.84 \times 10^{-12} \times 12}{2 \times 10^{-6}} = 5.3 \times 10^{-13}\text{ F}$$

The value of τ_{RC} is then $R_L C_j$ or $50 \times 5.3 \times 10^{-13} = 2.65 \times 10^{-11}\text{ s}$.

The total response time for the detector τ may then be obtained from eq. (7.36):

$$\tau = [(2 \times 10^{-11})^2 + (2.65 \times 10^{-11})^2]^{1/2} = 3.32 \times 10^{-11}\text{ s}$$

Although the above two effects are the main causes of finite response times in p-i-n photodiodes, there are others. One that should be mentioned arises from carrier diffusion: carriers generated outside the high field regions may contribute to the output provided the electron-hole pairs manage to diffuse into the high field region before recombination takes place. The problem with this is that carrier diffusion is inherently a relatively slow process and thus can give rise to a delayed response by the photodiode. The time taken for excess carriers to diffuse a distance d may be written (ref. 7.7)

$$\tau_{\text{diff}} = \frac{d^2}{2D_c} \quad (7.37)$$

where D_c is the minority diffusion coefficient. Thus distances of the order of micrometres typically involve carrier diffusion times of the order of nanoseconds (Example 7.8) and so carriers generated outside the depletion region may thus cause a 'slow' tail to be present in the response as illustrated in Fig. 7.32. Normally in a p-i-n photodiode the number of such carriers is sufficiently small that they cause few problems.

EXAMPLE 7.8 Diffusion time of carriers in Si

We consider the time taken for electrons to diffuse through a layer of p-type silicon $5\text{ }\mu\text{m}$ thick. Taking $D = 3.4 \times 10^{-3}\text{ m}^2\text{ s}^{-1}$, we obtain from eq. (7.37) that

$$\tau_{\text{diff}} = \frac{(5 \times 10^{-6})^2}{2 \times 3.4 \times 10^{-3}} = 3.7 \times 10^{-9}\text{ s}$$

7.3.6.5 Schottky photodiodes

In a *Schottky* photodiode a metal-semiconductor junction (section 2.8.6.2) takes the place

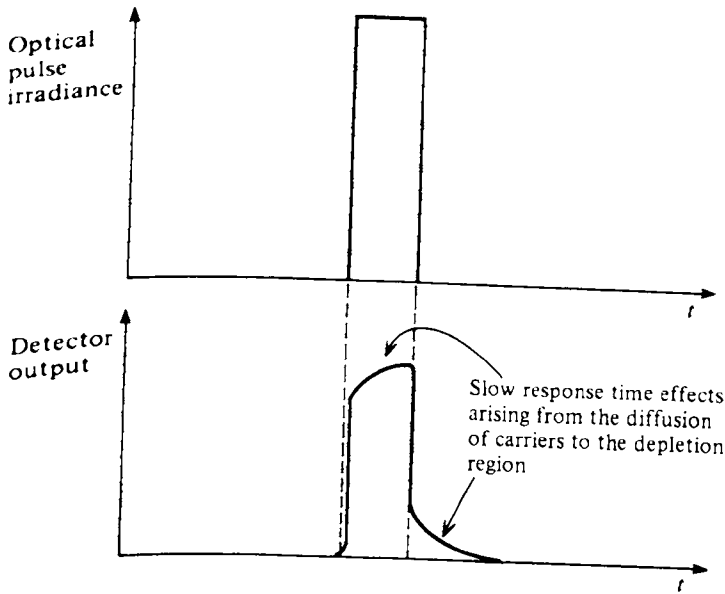


FIG. 7.32 Response of a p-n junction detector to a narrow optical pulse. The effects arising from the relatively long time it takes for carriers generated away from the depletion region to diffuse to the junction are shown.

of the semiconductor-semiconductor junction in the p-n photodiode. A thin (≤ 20 nm) metal coating (usually gold) is applied to a suitable semiconductor surface, as shown in Fig. 7.33(a). To improve transmission through the gold layer an antireflection coating is usually also applied. The energy band structure in the region of the junction is shown in Fig. 7.33(b). It can be seen from this that when an electron-hole pair is generated within the depletion region, the electron and hole will be separated by the action of the internal field just as in the p-n

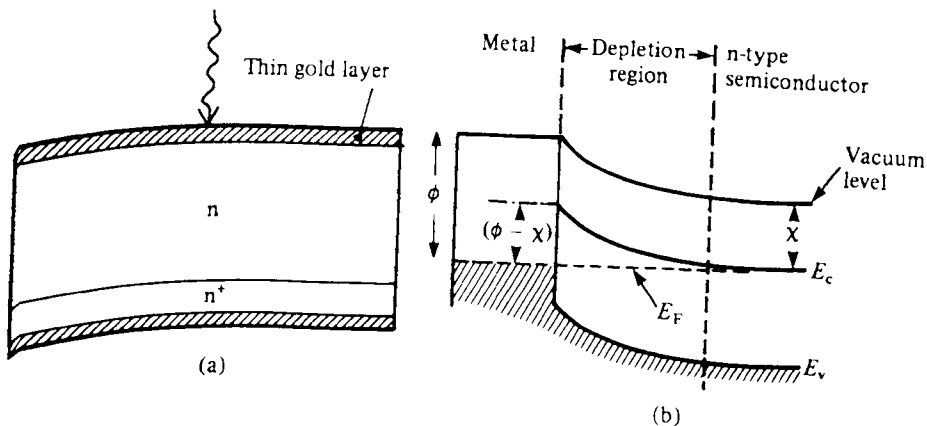


FIG. 7.33 Basic structure of a Schottky photodiode (a) and the energy level behaviour in the region of the junction (b). A potential barrier of height $\phi - \chi$ is formed between the metal and semiconductor.

junction photodiode. The main advantage of the Schottky photodiode is at wavelengths appreciably shorter than the bandgap wavelength λ_g . Here, in a normal p–n junction diode heavy absorption in the semiconductor surface layer gives rise to a much reduced quantum efficiency. In the Schottky diode the surface antireflection coating–metal surface layer can be much more highly transmitting at these wavelengths leading to higher efficiencies. However, the devices are relatively difficult to fabricate (great care is needed, for example, in obtaining the correct thicknesses of the antireflection coating and metal layer) and offer no significant advantages over p–n junction devices at wavelengths closer to the bandgap wavelength.

Schottky photodiodes are available using a variety of semiconductors such as Si, GaAs, GaP and GaAsP and are mainly used for their relatively high blue and UV sensitivity. It is of interest to note that the device can respond to photons that have a lower energy than that of the semiconductor bandgap. An inspection of Fig. 7.33(b) shows that photons which are absorbed in the metal and which have energies as low as $e(\phi - \chi)$ can cause electrons to surmount the metal–semiconductor barrier by thermionic emission and contribute to the photocurrent.

Photodetectors have also been made using metal–semiconductor–metal (MSM) junctions. These can be made very simply by forming two interdigitated metal contacts on top of a fairly thin (say 3 μm) undoped semiconductor layer (Fig. 7.34). The spacing and area of the contacts is such that at fairly modest bias voltages the semiconductor layer under the electrodes is almost completely depleted whilst ensuring that the light-sensitive area presented to incoming radiation is between 50% and 75% of the total area. The energy band diagram is shown in Fig. 7.35. The dark current is mainly determined by thermionic emission of both holes and electrons over the respective barriers and is comparable with that in p–i–n photodiodes (i.e. 1 nA or lower). It is interesting to note that the device capacitance is approximately four times smaller than a p–i–n junction of comparable light-sensitive area (only half the spacing between the electrodes is filled with dielectric, the other half is filled with air). This low capacitance is obviously very helpful in achieving a high speed of response. The spacing between the metal contacts can be made smaller than for a vertical diode, giving shorter carrier transit times and again increasing response speeds. Bandwidths of 20–50 GHz are possible.

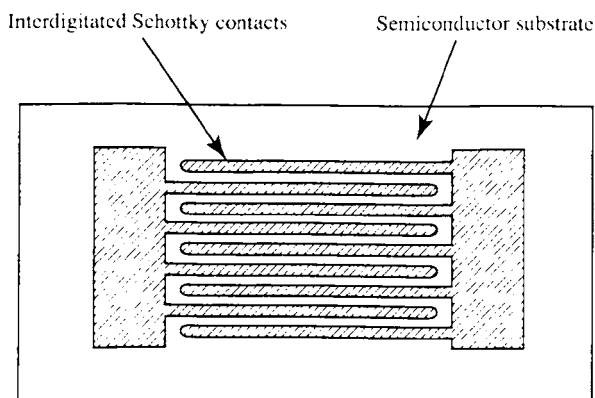


FIG. 7.34 The interdigitated electrode structure used for MSM photodiodes.

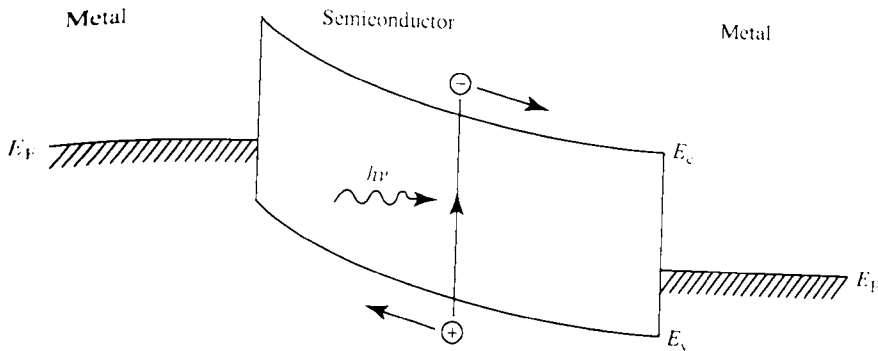


FIG. 7.35 Energy band diagram of a metal-semiconductor-metal junction under bias.

Such devices have been successfully made using GaAs as the semiconductor. As far as $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ is concerned there is a slight problem in that the Schottky barrier height is quite small which results in a high dark current. This may be reduced by interposing a thin $\text{In}_{0.52}\text{Al}_{0.48}\text{As}$ surface layer between the contacts which enhances the barrier height. These devices hold considerable promise as they are relatively easy to fabricate and also to integrate with amplifier structures (see section 9.4).

7.3.6.6 Avalanche photodiodes

As explained above, most fast photodiodes are designed for use with a $50\ \Omega$ load impedance and the voltage output ($= i_x R_L$) often requires considerable amplification. Useful internal amplification of the photocurrent is achieved in the avalanche photodiode (APD). In this device, a basic p-n structure is operated under very high reverse bias. Carriers traversing the depletion region therefore gain sufficient energy to enable further carriers to be excited across the energy gap by impact excitation. The process is illustrated in Fig. 7.36.

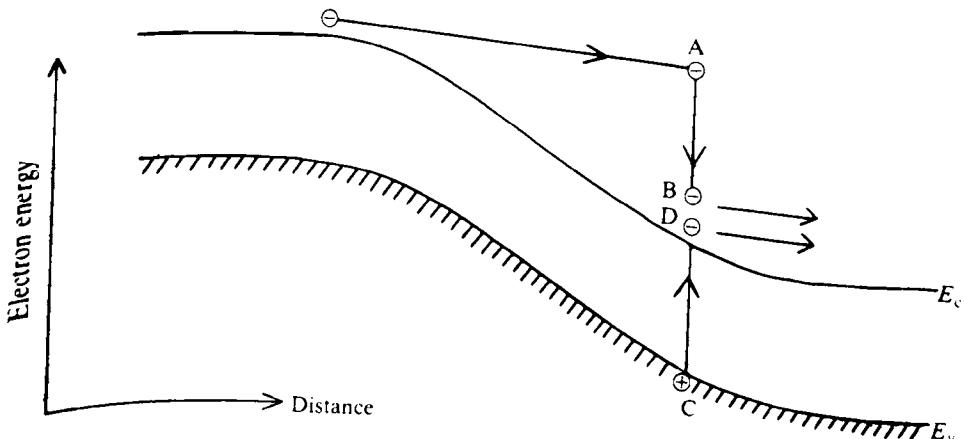


FIG. 7.36 Illustration of the principle of operation of an avalanche photodiode. An electron having reached the point A has sufficient energy above the conduction band bottom to enable it to excite an electron from the valence band into the conduction band (C → D). In so doing it falls from A to B.

An electron having reached point A on the diagram has enough energy above the conduction band bottom such that it can collide with an electron from the valence band and raise it to the conduction band ($C \rightarrow D$) (the minimum energy required to initiate this process was discussed in section 4.3 when dealing with cathodoluminescence). This generates a new electron-hole pair; in so doing, the electron will of course lose an equivalent amount of energy and move from A to B. The newly generated carriers may both subsequently generate further electron-hole pairs by the same process. The probability that an ionizing event takes place is governed by the carrier densities and also the *ionization coefficients*. The latter vary rapidly with electric field \mathcal{E} according to a relation of the form $\exp(-A/\mathcal{E})$ where A is a constant. In addition it should be noted that the magnitudes of the ionization coefficients can differ depending on the type of carrier. In silicon, for example, the ionization coefficient for electrons is considerably greater than that for holes, whilst in germanium the two are almost identical.

Current gains in excess of 100 are readily achievable. However, as shown in Fig. 7.37, the current gain is very sensitive to the value of the bias voltage, and if the bias voltage is made too large there is the danger of creating a self-sustaining avalanche current that flows in the absence of any photoexcitation, which sets an upper limit on the voltage that may be used. In fact non-uniformities within the device can cause small regions of premature breakdown known as *microplasmas*. Obviously the device requires a very stable voltage supply and, in addition, since the avalanche process itself depends on temperature, temperature stabilization is required. Another problem that can occur is that of excessive leakage currents at the junction edges. In silicon APDs this latter problem can be addressed by the use of a *guard ring* as shown in Fig. 7.38. This also serves to restrict the avalanche region to the central illuminated part of the cell and thus helps to reduce premature breakdown. Unfortunately the guard ring also increases the capacitance of the device thus restricting the high frequency performance.

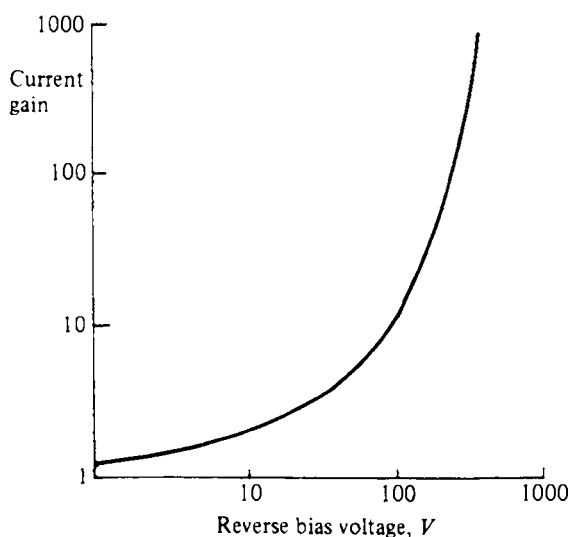


FIG. 7.37 Typical variation of current gain with reverse bias voltage for an avalanche photodiode.

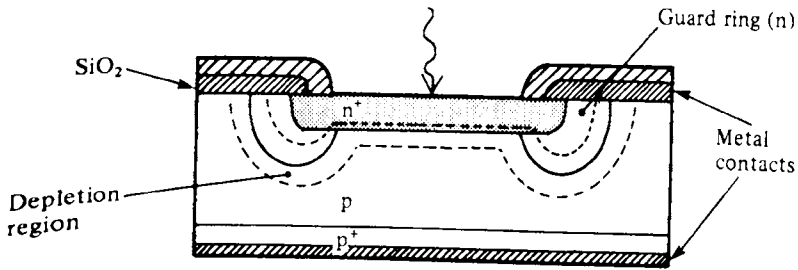


FIG. 7.38 Silicon avalanche photodiode with guard-ring structure. The guard ring is a region of comparatively low doping, and hence the depletion region extends an appreciable distance into it. Thus, in the vicinity of the guard ring the total depletion layer thickness is greater, and hence the maximum electric field strengths are less, than in the central region.

Another type of avalanche device is the *reach-through avalanche photodiode* (RAPD). This has a $p^+ \pi p n^+$ structure, as shown in Fig. 7.39(a) (π indicates lightly doped p material). When a sufficiently high external voltage is applied the field distribution is as shown in Fig. 7.39(b). The field extends all the way through (i.e. it *reaches through*) the π region to the $p n^+$ junction where the fields are high enough to give rise to avalanche gain. In operation incident radiation is absorbed in the relatively wide π region. The field then separates the charge carriers with the electrons moving to the avalanche region where they take part in the avalanche process as described above. Since electrons have a much larger ionization coefficient than holes in silicon, the fact that only electrons initiate the avalanche process has little bearing on the overall current gains achieved. In addition, because the applied voltage is 'shared' between the (wide) π region and the avalanche region the gain becomes relatively insensitive to variations in applied voltage.

As in the case of photodiodes, the extension of the range of APDs up to a wavelength

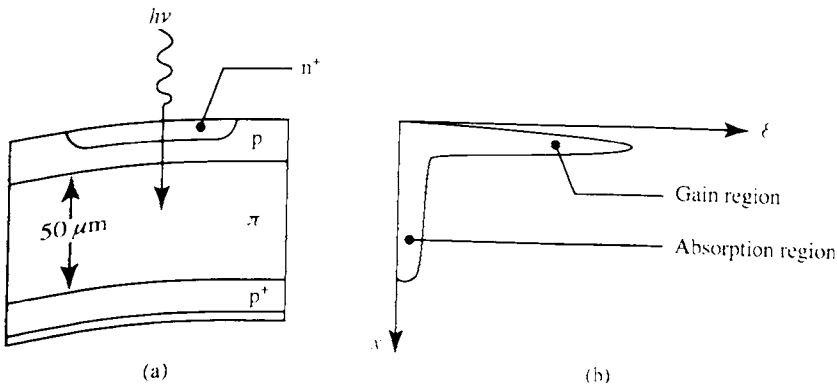


FIG. 7.39 (a) Structure of a silicon RAPD device. (b) The field distribution across the RAPD; the high field gain region is across the $p-n^+$ junction.

of $1.6\ \mu\text{m}$ requires the use of semiconductors with smaller bandgaps than silicon. One immediate problem with such materials is that dark currents will be larger and although cooling can help, this obviously adds to the complexity of the detector. Germanium has been used reasonably successfully, but the fact that the electron and hole ionization coefficients are almost equal gives rise to relatively large amounts of noise arising from variations in the avalanche gain process itself (see the comments at the end of this section). As with junction diodes considerable effort has been put into developing InP/InGaAs devices. The problem of increasing dark currents with decreasing bandgaps has been solved by separating the absorption and gain regions (as in the RAPD) with the absorption occurring in the narrow bandgap material $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ whilst the avalanche gain occurs in the wider bandgap material InP. This structure is called the *SAM* (Separate Absorption and Multiplication) *APD* and is illustrated in Fig. 7.40.

NOISE CONSIDERATIONS

As well as the usual sources of noise discussed previously another factor appears with the avalanche photodiode. The exact number of carrier multiplication events produced by a single charge carrier as it moves through the high field region will be subject to statistical variation (because the distance travelled before an ionizing event occurs will be subject to statistical variation). If there were no such fluctuation in the gain process then the shot noise contribution would simply be multiplied by the gain M . However, because of the variation in M we may write

$$\Delta i_{\lambda} = M[2F(M)(i_{\lambda} + i_D)e\Delta f]^{1/2} \quad (7.38)$$

where i_{λ} is the photogenerated current, i_D the dark current and $F(M)$ a function called the *excess noise factor*. Where electrons alone are injected into the high field region it has been

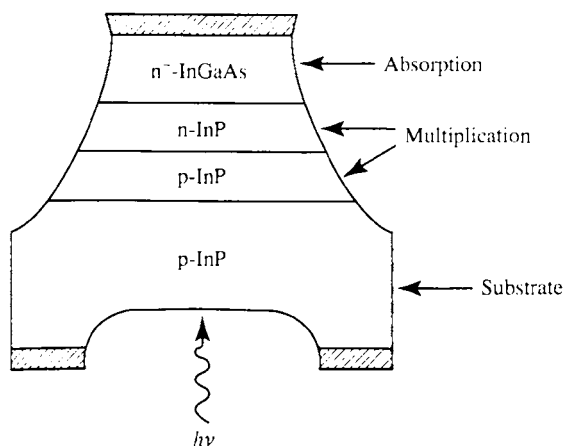


FIG. 7.40 Basic structure of a SAM APD device based on InP. Note that in this example the radiation is incident through a window that has been etched into the substrate.

shown (ref. 7.8) that $F(M)$ can be written as

$$F(M) = M \left[1 - \left(1 - \frac{1}{r} \right) \left(\frac{M-1}{M} \right)^2 \right] \quad (7.38a)$$

where r is the ratio of electron to hole ionization coefficients (for the case of hole injection the factor r in eq (7.38a) is replaced by $1/r$). A semiempirical expression that is sometimes used for $F(M)$ is M^x , with x taking a value between 0 and 1 depending on the material.

An inspection of eq. (7.38a) shows that when $r \rightarrow \infty$ then $F(M) \rightarrow 2 - 1/M$, whereas when $r \rightarrow 1$ then $F(M) \rightarrow M$. It is thus advantageous from a noise point of view to use materials where the ionization probabilities of the two types of carrier are very different from each other.

7.3.6.7 Phototransistors

The *phototransistor* is another device, like the avalanche photodiode, where the current flow from a p-n junction detector is internally amplified. The construction is basically that of a junction transistor, with the base region exposed to the incident radiation. Normally no external connection is made to the base (see Fig. 7.41a). To understand the operation of the device, we consider the external currents to be as shown in Fig. 7.41(b). The base current i_b will be supplied by the photogenerated current.

We must have

$$i_c = i_e - i_b$$

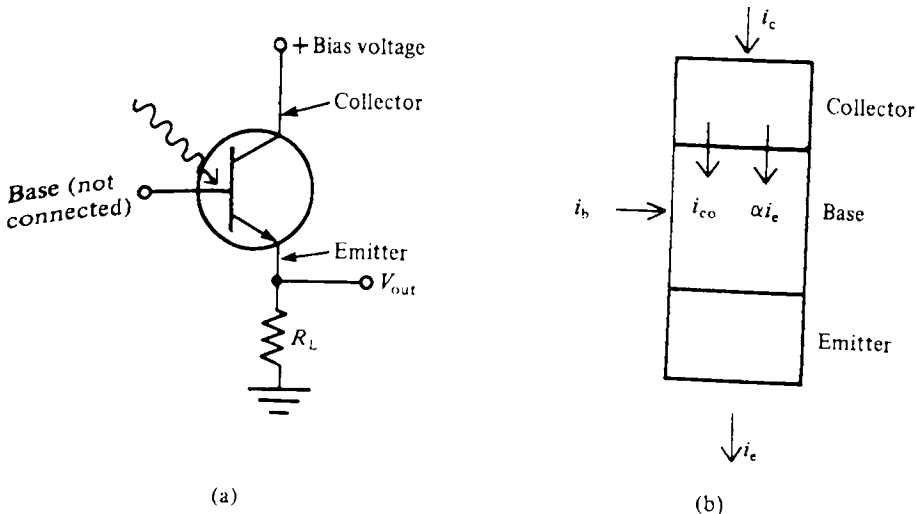


FIG. 7.41 External connections made to an n-p-n phototransistor (a). Light absorbed within the base region causes an emitter current to flow through the load resistor R_L and thus a signal voltage will appear across it. (b) The currents assumed to be flowing in the phototransistor.

where i_c and i_e are the collector and emitter currents respectively. The collector current has two components: (a) the normal diode reverse saturation current i_{co} , and (b) that part of the emitter current that manages to cross to the collector. (The current is carried by minority carrier diffusion across the base, and not all the minority carriers leaving the emitter will reach the collector.) We write this latter current as αi_e , where α is slightly less than unity (α is known as the *common base current gain*). Thus

$$i_{co} + \alpha i_e = i_c - i_b$$

whence

$$\begin{aligned} i_c &= \frac{i_b + i_{co}}{(1 - \alpha)} \\ &= (i_b + i_{co}) \left(\frac{\alpha}{1 - \alpha} + 1 \right) \\ &= (i_b + i_{co})(h_{fe} + 1) \end{aligned}$$

where $h_{fe} = \alpha/(1 - \alpha)$ is known as the *common emitter current gain* of the transistor. Typical dark values for h_{fe} in phototransistors are about 100. With no incident radiation $i_b = 0$ and the current flowing, $i_{co}(h_{fe} + 1)$, is the dark current of the device. This is obviously larger than for comparable p-n junction devices when the dark current in this notation is just i_{co} .

When illuminated there will be a base current of magnitude i_λ , where, from eq. (7.25), $i_\lambda = \eta(I_0 A e \lambda_0 / hc)$. The external current flowing is now $(i_\lambda + i_{co}) \times (1 + h_{fe})$ which, if $i_\lambda \gg i_{co}$, is equal to $i_\lambda(1 + h_{fe})$. Thus the device gives us internal *gain*, and has a responsivity lying between that of a p-i-n photodiode and an avalanche photodiode.

Silicon-based phototransistors are readily available at low cost; a typical device structure is shown in Fig. 7.42. Such detectors usually suffer from a poor frequency bandwidth, often being limited to a few hundred kilohertz. This arises both from the high capacitance of the base-collector junction and the long carrier transit times across the base region. However, the presence of internal gain can greatly simplify detection circuitry where the small bandwidth is not a problem (e.g. in remote control devices for TVs and videos).

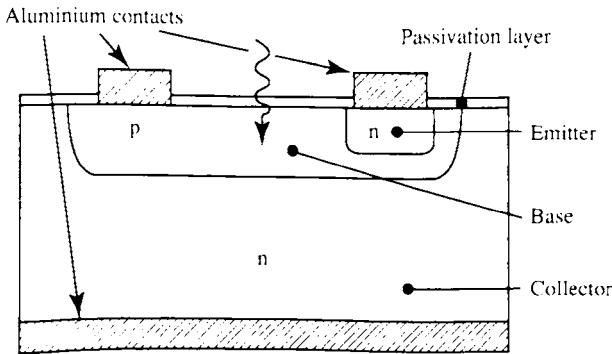


FIG. 7.42 Structure of a silicon phototransistor.

It is possible to make phototransistors with much higher frequency bandwidths. Figure 7.43, for example, shows a structure for an n-p-n phototransistor based on an InGaAsP/InP heterojunction. As in the InGaAs/InP p-i-n photodiode of Fig. 7.31, incident radiation with a wavelength greater than $0.92\text{ }\mu\text{m}$ will pass unattenuated through the upper InP layer and wavelengths up to the bandgap wavelength of the base material will then be absorbed in the base.

7.3.6.8 Intensified photodiodes

A recently developed device that is a hybrid between a photomultiplier and a photodiode is the *intensified photodiode*. In this device, which is illustrated in Fig. 7.44, the incoming

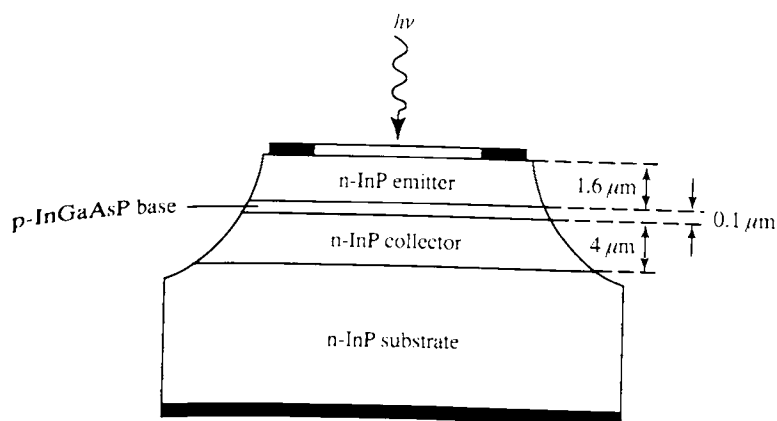


FIG. 7.43 Structure of an n-p-n phototransistor based on an InGaAsP/InP heterojunction.

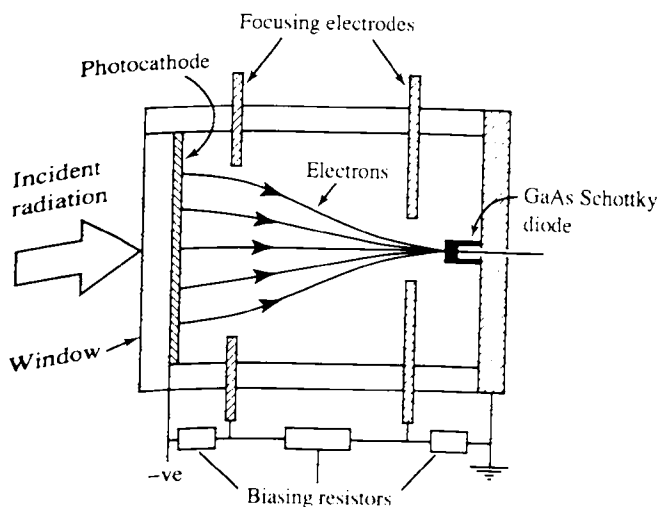


FIG. 7.44 The intensified photodiode. The voltage for the focusing electrodes is applied via a potential divider circuit similar to that used in a photomultiplier.

radiation strikes a photocathode which is situated on the inside of a vacuum tube as in a photomultiplier. The emitted photoelectrons are then accelerated down the tube (to an energy of about 8 keV) and focused onto a small area semiconductor p–i–n junction or Schottky diode. As the high energy electrons travel through the semiconductor material each generates several thousand secondary electrons, thereby producing a current gain. Commonly used photocathode materials include GaAs and GaAsP which can cover the wavelength range from 340 nm to 830 nm.

In terms of performance their D^* values can be appreciably larger than comparable photomultipliers or avalanche photodiodes. One drawback is that they require a relatively high operating voltage (8–15 keV). Usually the intensified photodiode is supplied with a compact high voltage supply operating from a 12 V supply.

7.3.7 Detector arrays

In section 7.3.5.3 we have seen how an optical image may be converted into an electrical signal by using a vidicon-type device. An alternative approach is to use an array of discrete detectors. The image may be focused directly onto a two-dimensional array, or alternatively the image may be scanned past a one-dimensional array. The problem with either of these more direct approaches is that to obtain a scan of the image requires that the output of each detector must be read out sequentially. Given that an array may contain many thousands of detectors, the problem of correctly interconnecting them is not trivial. Although nowadays this is not a real problem, it was in the 1970s and so an alternative approach was proposed by researchers at Bell Laboratories. This involved using a *charge-coupled device* (CCD) structure. The basic building block of this is the *metal–oxide–semiconductor* (MOS) capacitor. This is formed by growing a layer of silicon dioxide (SiO_2) on a p-type silicon substrate; a metal electrode is then evaporated on top of the oxide layer (Fig. 7.45). The metal electrode is known as the *gate* and is biased positively with respect to the silicon. Photogenerated electron–hole pairs within the silicon will be separated, with the electrons

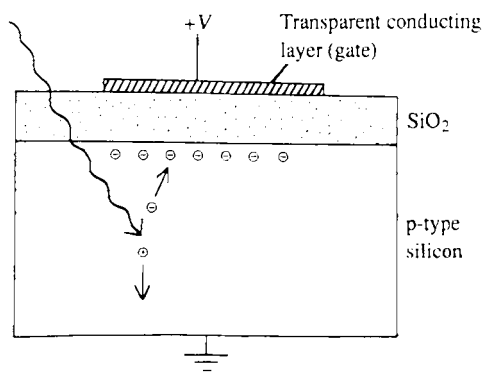


FIG. 7.45 MOS capacitor structure. When the gate is biased positively, photogenerated electron–hole pairs become separated and the electrons then become trapped at the SiO_2 –Si boundary beneath the gate electrode.

being attracted to the surface of the silicon under the gate, where they will remain whilst the gate voltage is positive. The electrons are effectively trapped within a potential well formed under the gate contact. The amount of trapped charge will be proportional to the total integrated light flux falling onto the device during the measurement period.

The problem now is to 'read out' this charge sequentially along a line of such detectors. This can be done by passing the charge from detector to detector. There are several ways of achieving this, the basic idea being illustrated in Fig. 7.46. The gate potentials are supplied from three voltage lines (L_1, L_2, L_3) each one being connected to every third electrode (G_1, G_2, G_3). We suppose that initially the potential of L_1 is at some positive value V_g , whilst L_2 and L_3 are at zero potential. Photogenerated charge will be trapped under the G_1 electrodes in proportion to the amount of light falling on these elements (Fig. 7.46b). After a suitable integration time, the charge may be moved along the chain of MOS capacitors by applying a repeated sequence of potentials to the gate supply lines. Thus, suppose we apply a voltage V_g to L_2 whilst maintaining L_1 at V_g ; the charge initially under L_1 will now be shared between G_1 and G_2 (Fig. 7.46c). Next we reduce the potential of L_1 to zero. All the charge that was initially under G_1 is then under G_2 (Fig. 7.46d). Continuing this cycle will progressively move the charge along the line of MOS capacitors from left to right. At the end of the scan of the ' G_1 ' detector outputs.

For obvious reasons, this is known as a *three-phase* scheme. In practice, several other schemes are possible and ref. 7.9 may be consulted for more details. One problem with the

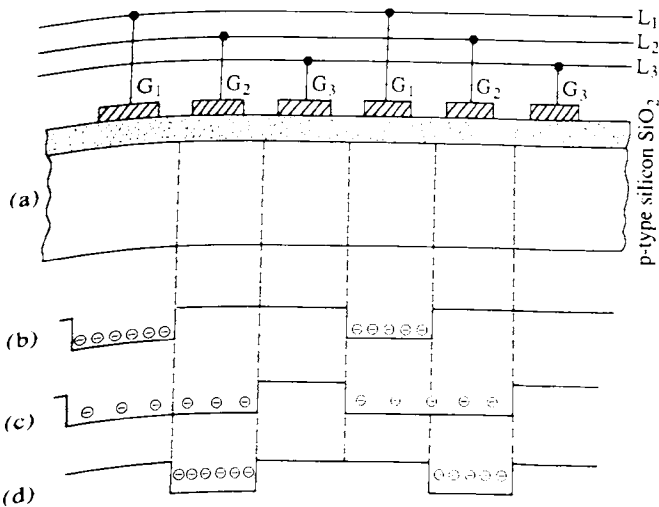


FIG. 7.46 Basic CCD array composed of a line of MOS capacitors (a). The initial charge distribution within the potential wells when G_1 has a positive voltage applied ($= V_g$) and $G_2 = G_3 = 0$ is shown in (b). In (c) $G_1 = G_2 = V_g$, $G_3 = 0$ and the charge has now spread out under both G_1 and G_2 . In (d), $G_1 = 0$, $G_2 = V_g$ and $G_3 = 0$. The charge that initially was under G_1 has now moved to the right to be under G_2 . Note that no charge should be generated under the gates G_2 and G_3 ; these elements are therefore screened from incident light.

device as described here is that a new light scan cannot be carried out until all of the charge has been transferred along the array. A faster scanning rate may be achieved with the layout indicated in Fig. 7.47. Here, a second CCD array (the *transport register*), which is shielded from incident radiation, lies alongside the first. Once a charge image has been built up in the sensing array it is then transferred 'sideways' to the transport register and can be read out sequentially along the transport register. This readout takes place at the same time as a new image is being built up.

Two-dimensional arrays based on the above one-dimensional designs are also possible, and are known as *frame transfer* devices. Here the transfer registers feed into a *readout register* running down the edge of the device (Fig. 7.48a). The contents of each line are read out in sequence into the readout register so that the signal appearing at the end of the register represents a line-by-line scan of the image. A problem with this scheme is that the presence of the transfer register reduces the resolution possible by increasing the distance between the sensing arrays. It is possible to dispense with the transfer register altogether by adopting the design shown in Fig. 7.48(b). Here, each array is made twice as long as before, with the second half of the array being shielded from the incident radiation. Then, after a charge image has been built up it is moved along each array into the shielded section where it can be stored until it can be transferred into the readout register as before. Again, readout takes place whilst a new image is being built up. This scheme is known as *interline transfer*. Although it offers better resolution than frame transfer, it is also somewhat slower because of the relatively long time taken to transfer charge along the sensing array and into the storage section.

It will be evident that whichever of these schemes is adopted the photogenerated charge has to undergo a large number of transfers before being 'read off'. It is absolutely essential, therefore, that as little charge as possible gets 'lost' on the way. The fraction of the total charge which is successfully moved in each transfer is called the *charge transfer efficiency* (η_{ct}). Thus after m transfers the original charge will be reduced by a factor $(\eta_{ct})^m$. For the readout signal to provide a faithful representation of the pixel signal it is essential that η_{ct} is very close to unity. Thus consider a 1024×1024 CCD array which operates with $\eta_{ct} = 0.9999$ (i.e. 1 electron lost in 10 000). The maximum number of charge transfers will be 3×2048 (assuming a three-phase scheme) and the amount of the original charge that is transferred will be

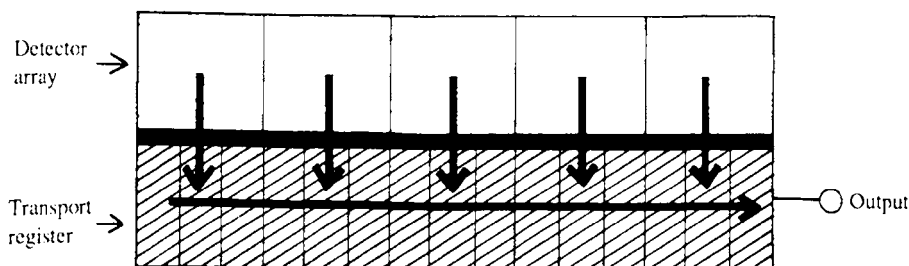


FIG. 7.47 Basic arrangement of the detector array and transport register in a linear CCD optical sensing array. The directions of charge transport are indicated by arrows and shading indicates those areas that are shielded from incident radiation.

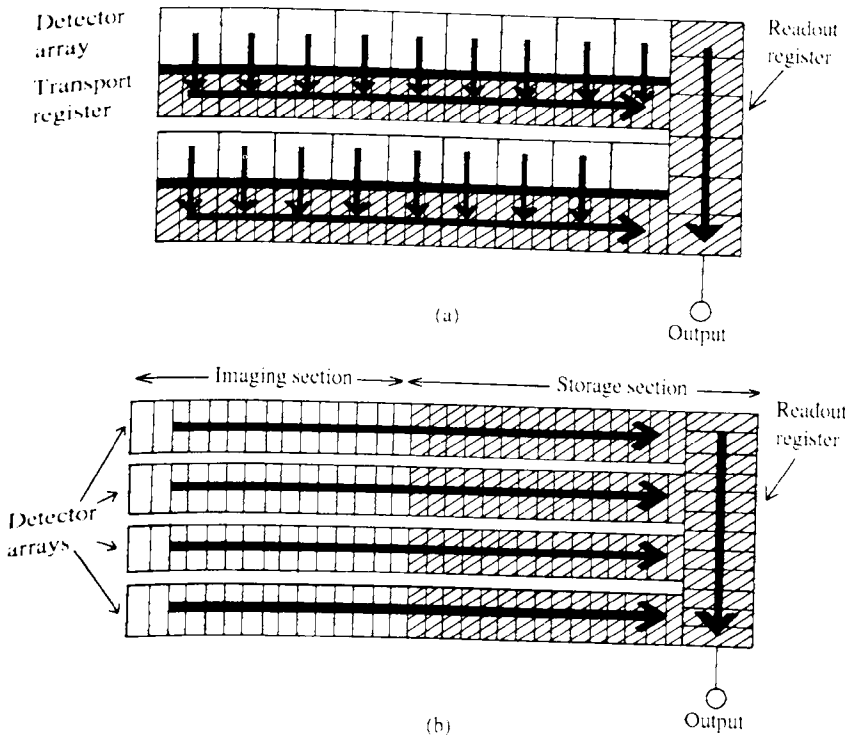


FIG. 7.48 Two schemes for reading out photogenerated charge from a two-dimensional CCD image sensor: frame transfer (a) and interline transfer (b). As in Fig. 7.47, charge flow is indicated by arrows and areas shielded from radiation by shading.

0.9999⁶¹⁴⁴ or 0.54. In current CCDs values for η_{ct} can reach 0.999 999. Such values have come about as a result of the virtual elimination of crystal defects in the materials used. When CCD devices are used in an environment such as space where there is a strong possibility of damage due to radiation bombardment then their performance can be severely limited.

In modern designs the gates are made from polysilicon (Fig. 7.49) which is reasonably transparent to radiation between 400 nm and 1100 nm. However, quantum efficiencies are then limited to about 0.35. The wavelength range can be extended downwards towards the UV by coating the front surface of the array with an appropriate phosphor material. Another approach is to thin the substrate sufficiently to allow light to be incident from below. Much higher quantum efficiencies (up to 0.9) are then possible and the short wavelength range is extended down to 200 nm. This procedure is not without its problems: the thinning process can cause the introduction of defects, and the very thin structure is difficult to support and to keep optically flat.

Two-dimensional arrays typically contain 1024×1024 pixels, although arrays with larger numbers are available. The pixel side lengths can vary between $7 \mu\text{m}$ and $50 \mu\text{m}$. The rate at which the images can be read out can vary greatly depending on the size of the array and the application. For example, when arrays are used in camcorders, video rates (i.e. 30 images

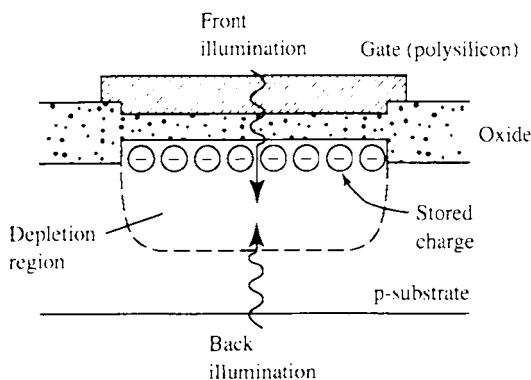


FIG. 7.49 Illustration of the polysilicon gate MOS structure in a CCD array; illumination may be either through the gate or from the back.

per second) are required although image quality is not as high as TV standards. The smallest light signals that can be detected depend on the number of 'dark' electrons; this is the number of electrons that accumulate under the gates during the integration time which are not photogenerated. At video rates of image transfer this can be about 10 electrons per pixel, although lower values can be obtained especially on cooling. Thus it is possible to detect light levels resulting in something like 40 photons per pixel. If yet higher sensitivities are required then the CCD array may be placed after an image intensifier.

Arrays of other types of photodetector are also available, one of the most common being based on the silicon photodiode. IR night vision devices can make use of detector arrays based on, for example, GaAs/GaAlAs, InSb and HgCdTe. The two latter detectors need to be cryogenically cooled. Another interesting development has been that of the *microbolometer*. Using a combination of photolithography and selective etching it has been possible to manufacture arrays of small silicon 'plates' $0.5\ \mu\text{m}$ thick and of area $50\ \mu\text{m} \times 50\ \mu\text{m}$ which are supported clear of the underlying silicon on two 'legs'. Incident radiation causes the temperature of the silicon slab to increase, and this temperature increase may then be detected by any of a number of means, for example by measuring the resistance of the slab (see section 7.2). Such detectors can have comparable performance with the cooled arrays.

7.3.8 Liquid crystal light valves

The liquid crystal light valve is an optical to optical image transducer that takes a low irradiance light image and converts it into an output image which uses light from another source (and which, in consequence, may be of a much higher irradiance, e.g. a laser beam). Although not primarily used for radiation detection purposes the operation of the device is more conveniently dealt with in this chapter than elsewhere. It was first developed by the Hughes Aircraft Company in the 1970s and has undergone substantial development since. Figure 7.50 shows the basic construction of the first-generation light valve. The device consists of several layers at the heart of which is a nematic liquid crystal cell. The 'read' beam first

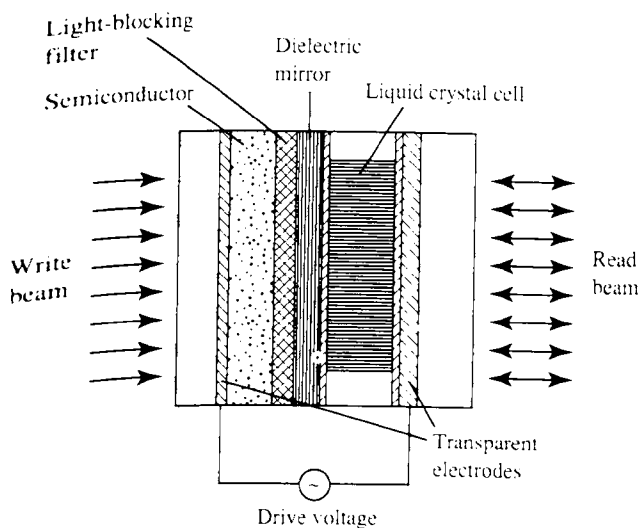


FIG. 7.50 Structure of a first-generation liquid crystal light valve.

traverses the liquid crystal layer, is reflected from a dielectric mirror, repasses through the liquid crystal layer and then exits from the device. The exact workings of the liquid crystal cell will be dealt with later since it operates in a slightly different mode from that described in section 4.8. For the moment we assume that the liquid crystal cell-mirror combination can be switched from a non-reflective mode to a reflective mode by the application of a suitable voltage. When no writing light is present the cadmium sulfide photoconductive layer has a very high resistance so that the potential, which is applied across the whole device, is mainly dropped across this layer rather than the liquid crystal, and consequently the liquid crystal cell-mirror combination is non-reflecting. When a writing beam is present radiation is absorbed within the cadmium sulfide layer and its resistance falls; the potential across the liquid crystal then increases causing the 'read' beam to be reflected.

If the liquid crystal cell were operated in the 'normal' way (e.g. as in Fig. 4.26) it would be necessary to have a polarizer between the cell and the dielectric mirror, which has a high resistance and thus would always have a considerable potential drop across it. In the *hybrid field effect mode* the liquid crystal molecules have a twist angle of only 45° (in contrast to the more usual 90°). The required polarizer and (crossed polaroid) analyzer functions are combined at the 'read' input by using a polarization beam splitter (Fig. 7.51). Light from the source is polarized as it is reflected into the device, whilst for the returning beam only light with an orthogonal polarization will be transmitted straight through the beam splitter. With no potential applied across the liquid crystal cell the polarized input 'read' radiation undergoes a 45° twist as it passes through the cell; it is then reflected and 'untwists' on its way back to give exactly the same state of polarization it had on entering. However, the analyzer will only transmit the orthogonal polarization, and so the radiation is blocked. As soon as a potential is applied to the cell, however, the molecules start to move away from their uniform 45° twist directions. The radiation that is transmitted is then not completely linearly

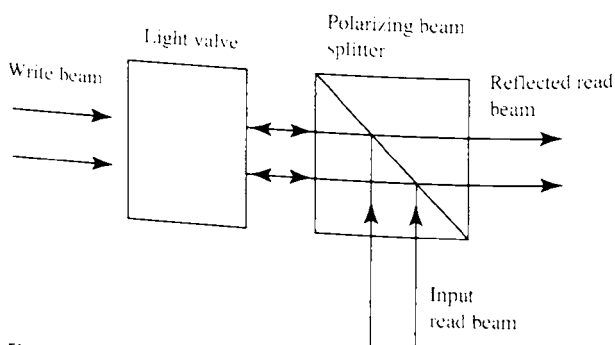


FIG. 7.51 Use of a polarization beam splitter to achieve the requirements for polarizer and analyzer at the input to the liquid crystal light valve.

polarized and thus some of it will be able to pass through the analyzer. With a sufficiently large potential applied, all the molecules line up along the field and again the output polarization is the same as the input, leading to the radiation again being blocked. The overall reflectivity of the cell as a function of applied voltage is shown in Fig. 7.52 (note that this characteristic is quite different from that of Fig. 4.27). Between the photoconductive layer and the dielectric mirror is a light-blocking layer; this is present to ensure that none of the 'reading' beam reaches the photoconductive layer.

One of the main problems with first-generation types is their relatively slow response time which arises from the presence of traps in the cadmium sulfide giving a long photoconductive response time. In second-generation types the cadmium sulfide layer is replaced by a silicon MOS structure as illustrated in Fig. 7.53. The device operates in a two-phase cycle; in the *depletion* (active) phase the applied voltage is such that the π -silicon is depleted of charge carriers. The electron-hole pairs that are generated by the input 'write' beam are separated, with the electrons moving to the Si/SiO₂ interface. When the charge arrives at the interface a grid of n-type 'microdiodes' serves to contain the charge within well-defined 'cells' and

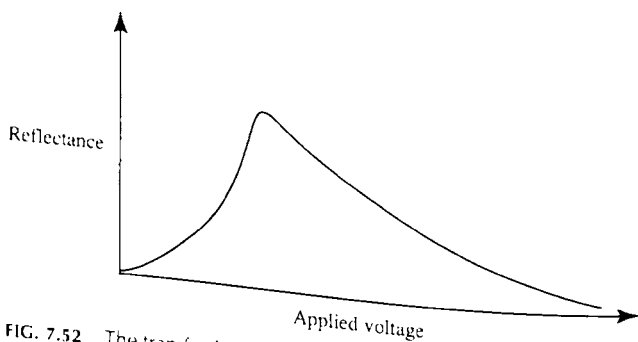


FIG. 7.52 The transfer function of a liquid crystal cell operated in the hybrid field effect mode with a 45° twist angle.

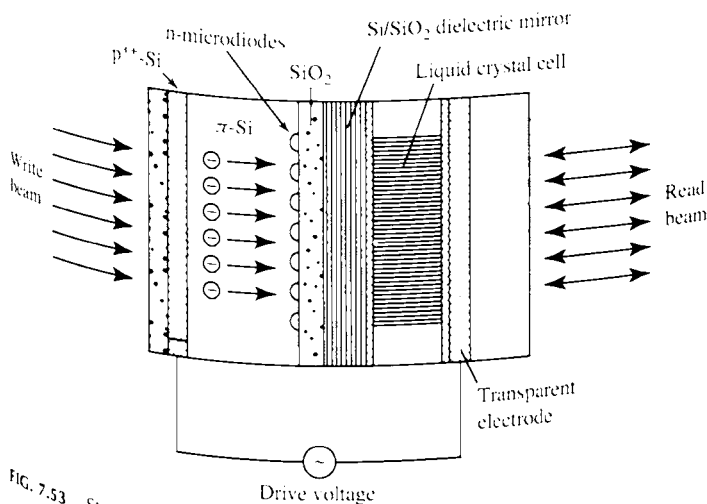


FIG. 7.53 Structure of a second-generation liquid crystal light valve.

prevents sideways diffusion and the consequent smearing of the charge pattern. The presence of this charge then activates the liquid crystal layer. (The exact way in which this occurs is in fact rather involved; ref. 7.10 should be consulted for further details.) The other operating phase is the *accumulation* phase, during which an opposite polarity pulse is applied; the electrons residing at the Si/SiO₂ interface are then attracted towards the back contact and recombine with the majority carriers in the bulk π material. This phase usually only occupies a few per cent of the total cycle time of the device. The dielectric mirror is conveniently made from alternating pairs of quarter-wave thicknesses of Si and SiO₂. The response times are now limited by that of the liquid crystal cell and not the photoconductor and are typically a few milliseconds. If colour images are required three light valves may be used with the input light being split up into the three primary colours, each one being sent to the appropriate light valve. The returning light may then be recombined into a single beam.

An interesting variant is the CCD-addressed device, in which the switching potentials for the liquid crystal cell are provided electrically via a CCD array structure formed on the back of the liquid crystal cell. The CCD array here acts in the opposite sense to that discussed in section 7.3.7; thus an electrical serial input is converted into a two-dimensional charge distribution over the back of the device. This charge is then transferred across a silicon layer onto the back of the mirror structure where it influences the reflection properties of the liquid crystal cell in exactly the same way as in the optically activated device. This enables an image to be produced directly from electrical information stored in a computer, for example.

Liquid crystal light valves have a number of applications, one of the most important being undoubtedly in large screen projection systems whereby the input is obtained from, say, a cathode ray tube, and the read beam is a powerful source such as an arc lamp. However, light valves can also be used in image processing and optical data processing systems (ref. 7.11).

7.3.9 Photon counting techniques

A useful technique that is sometimes used when dealing with very low level signals is *photon counting*. Usually the arrival of a single photon at a detector gives rise to a very small output signal. However, if the detector exhibits considerable internal gain (as in the photomultiplier and the avalanche photodiode) it may be possible to detect the arrival of a single photon as an output voltage pulse. If optical power P_λ at wavelength λ_0 is incident on such a detector whose quantum efficiency is η , then we would expect to see a pulse rate $R_p = \eta P_\lambda \lambda_0 / (hc)$. Ideally the pulses should all have the same amplitude, but the gain values will always exhibit a statistical fluctuation about a mean and consequently we should expect the output pulse heights to be similarly distributed.

Not all the output pulses will be due to the arrival of a photon as there are bound to be spurious noise pulses present. These may be allowed for in two ways. First if the signal is switched on for a given time and then switched off for the same length of time we would expect approximately the same number of noise pulses in both time intervals, and hence a simple subtraction of the pulse counts obtained over the two time periods should enable a better signal count rate to be obtained. Secondly the noise pulses are likely to have a much wider range of pulse heights than the signal pulses, for example Fig. 7.54 shows a pulse height distribution containing both signal and noise from a photomultiplier specifically designed for photon counting. By only counting those pulses which lie between the limits expected from the signal pulses then a number of noise pulses will be rejected. When avalanche photodiodes are used for photon counting, it is usual to cool them to reduce thermally generated noise pulses, and to operate them with a reverse bias at or just above that normally required for breakdown to ensure that the arrival of a single photon will generate the largest possible output pulse. However, a quenching circuit is also required to

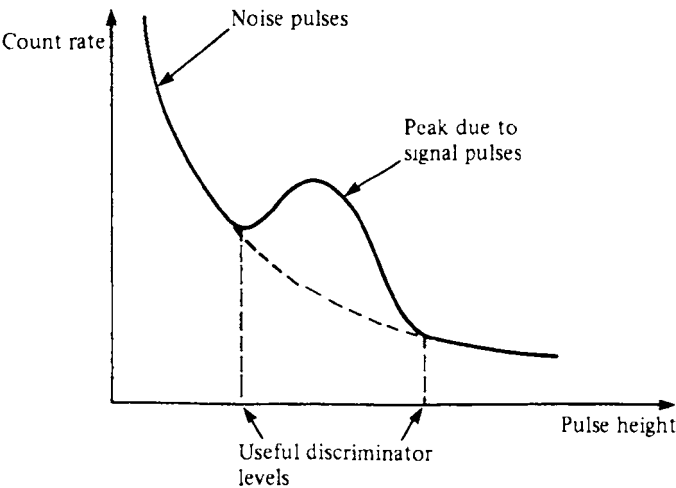


FIG. 7.54 Signal pulses have a relatively restricted range of pulse heights, whereas noise pulses show a much wider range of pulse heights. Discriminator levels can therefore be set to accept most of the signal pulses while a large proportion of the noise pulses will be rejected.

prevent the subsequent continuous avalanche breakdown that would normally follow such an event.

One problem with this scheme arises from the finite response time (τ_c) of the detection system. The output pulses will have a finite time width, and if the pulse rate (R_p) is too high then there is the danger that two pulses will arrive within a time equal to the response time and the two signal pulses will then merge into a single large output pulse and be either discounted or miscounted. This will happen when $\tau_c \approx 1/R_p$. This places an upper limit to the signal power that can be detected. An additional problem with the avalanche photodiode quenching circuitry is that there will be a period of time following a detected pulse when the detector is inactive.

7.3.10 Solar cells

One of the more obvious sources of renewable energy is the radiation from the sun. Just outside the earth's atmosphere the irradiance is about 1400 W m^{-2} . Because of attenuation in the atmosphere, this figure falls to between 500 W m^{-2} and 1000 W m^{-2} on the surface of the earth, depending on the angle of incidence of the radiation. The peak in the solar spectrum is at about $0.6 \mu\text{m}$ and it has a useful wavelength range of about $0.3 \mu\text{m}$ to $2 \mu\text{m}$. The aim of the solar cell is to convert as much of this radiation as possible into electrical energy.

In essence the solar cell is simply a p-n junction detector operated under conditions such that it can deliver power into an external load. The full i - V characteristics of a p-n junction under illumination are sketched in Fig. 7.55. The solar cell operates in the quadrant where i

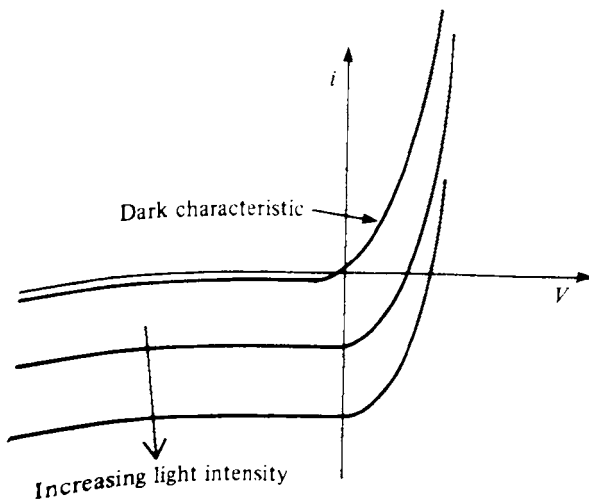


FIG. 7.55 Current-voltage characteristics of a p-n junction solar cell under various levels of illumination. The dark characteristic is that of an ordinary p-n junction diode. Under increasing levels of illumination the curve is progressively shifted downwards. The intercepts of the curves with the $V = 0$ axis give the values for the short circuit currents (i_{sc}), whilst the intercepts with the $i = 0$ axis give the values for the open circuit voltages (V_{oc}).

is negative and V is positive. Most power will be delivered when the product $i \times V$ is a maximum and this determines the optimum load impedance across the cell. It is fairly evident from Fig. 7.55 that we require both the open circuit voltage (V_{oc}) and the short circuit current (I_{sc}) to be as large as possible. From eq. (7.29) we see that the open circuit voltage is proportional to the factor $\ln[\eta I_0 e \lambda_c A / (h c i_0)]$, and thus we need to make i_0 , the reverse bias leakage current, as small as possible. Reference to eq. (2.51a) shows that this may be achieved by making the minority carrier populations (i.e. n_p and p_n) as small as possible. This implies that the doping on either side of the junction should be as large as possible. We saw in our analysis of the p-n junction photodetector that a signal could be obtained not only from photons that were absorbed within the depletion region but also from electrons and holes that had diffused to the depletion region after being generated outside it. In photodetectors this addition to the output was not regarded as advantageous since it increased the response time unduly. In the present instance, however, we are not at all concerned with response times and so the additional current flow is useful, especially as it is likely to arise from absorption of long wavelength photons where the absorption coefficient of the semiconductor may be relatively low. To increase this contribution the diffusion length of the carriers should be as large as possible. Since $L_c = (D_c \tau_c)^{1/2}$ this requires long minority carrier lifetimes. Unfortunately large values for the minority carrier lifetimes imply *low* values for the impurity doping concentrations (see eq. 4.16) and hence a compromise on the doping concentrations has to be reached. The most common structure for silicon solar cells, for example, is $n^+ - p^+$.

If the maximum power is obtained when the current is I_m and when the voltage is V_m then the *fill factor* F is defined by

$$F = \frac{V_m \times I_m}{V_{oc} \times I_{sc}} \quad (7.39)$$

In a good cell design, fill factors of as high as 0.8 can be obtained. The upper surface of a solar cell usually has a 'finger' network of ohmic contacts on it for efficient collection of the current. Typically these cover some 5–10% of the surface and their design needs some care to minimize the effective resistance that the electrons encounter as they travel from the junction to the external circuitry.

The first requirement of the solar cell material is that it has a bandgap which allows the absorption of a substantial part of the useful solar spectrum. The material must also exhibit long carrier diffusion lengths (in other words, high mobilities and long carrier lifetimes). As well as silicon there are several potential candidates within the II–VI and III–V groups of compounds, for example gallium arsenide, cadmium telluride and copper indium diselenide. Although these latter materials are in general more difficult to process than silicon they can be made into more varied and efficient structures (e.g. it is possible to have 'cascade' structures so that the incoming radiation encounters a series of cells with decreasing bandgaps which progressively remove longer and longer wavelengths). The expense involved in manufacturing such efficient structures can only be justified in space applications and in circumstances where the light is concentrated using some sort of lens or mirror structure so that only a relatively small area of solar cell is required.

At present, most commercially available cells are made from silicon which can be in single crystal, polycrystalline or amorphous form. Of these the single crystal cells are generally

the most efficient (10–12%) whilst amorphous cells, although cheaper to manufacture, have lower efficiencies (about 6%) and also suffer from light-induced degradation which reduces the efficiency, over a period of time, to some 80% of the initial value.

NOTES

1. In contrast to the situation in eqs (7.12) and (7.14), this expression for the r.m.s. noise explicitly involves f and hence we have to refer to the noise within the infinitesimally narrow frequency interval df .
2. Strictly speaking this relationship only applies if the system response functions as defined in Appendix 5 are both Gaussian; however, it is approximately valid for other pulse shapes.

PROBLEMS

- 7.1 A lead sulfide photodetector has a sensitive area of 10^{-4} nm^2 and is used to detect radiation of $2 \mu\text{m}$ wavelength. Using Fig. 7. 1, estimate the smallest signal power that it can detect if the detection system has a frequency bandwidth of 100 Hz.
- 7.2 Calculate the wavelength at which the energy of a photon becomes equal to the average thermal energy of atoms in a solid at room temperature.
- 7.3 Starting from eq. (7.1), that is

$$W = H \frac{d(\Delta T)}{dt} + G \Delta T$$

which represents the energy balance condition in a thermal detector element, and assuming that W and ΔT may be written in the form $W = W_0 + W_f \cos(2\pi f t)$ and $\Delta T = \Delta T_0 + \Delta T_f \cos(2\pi f t + \phi_f)$, show that:

- (a) $\Delta T_0 = W_0/G$; and
- (b) $\Delta T_f = W_f/(G^2 + 4\pi^2 f^2 H^2)^{1/2}$.

- 7.4 Show that the effective thermal conductance, G_R , for a radiative link at absolute temperature T may be written

$$G_R = 4\sigma A \epsilon T^3$$

where σ is Stefan's constant, ϵ the receiver surface emissivity and A its area. Hence show that the noise power fluctuations $(\Delta W_f)_R$, in a detector of bandwidth Δf limited by radiative exchange, is given by

$$(\Delta W_f)_R = 4(A\sigma\epsilon k T^5 \Delta f)^{1/2}.$$

Assuming $\epsilon = 1$, calculate $(\Delta W_f)_R$ for a detector of area 100 nm^2 and bandwidth 1 Hz at room temperature (300 K).

- 7.5 One fairly easily made bolometer consists of a long length of very fine metal wire (usually copper) bundled inside an insulating container (e.g. a thermos flask). This is

sometimes known as a ‘rat’s nest’ calorimeter. Show that the fractional change in the resistance of the wire (length L and radius r) when a heat pulse of energy H joules is input can be written

$$\frac{\Delta R}{R} = \frac{H\alpha}{\pi^2 L \rho S}$$

where ρ is the density of the metal, α its temperature coefficient of resistance and S its specific heat.

As mentioned in section 7.1.2 the change in resistance may be measured by incorporating the bolometer element into a Wheatstone bridge circuit (Fig. 7.4). Determine how the current through the galvanometer depends on the voltage applied across the bridge and the magnitude of the heat pulse (assume for simplicity that the resistances in the arms of the bridge are all equal before the heat pulse arrives). Hence determine the responsivity of such a device made from 100 m of fine copper wire given the following data for copper:

Density (ρ)	8930 kg m^{-3}
Specific heat (S)	$384 \text{ J K}^{-1} \text{ kg}^{-1}$
Resistivity	$1.7 \times 10^{-8} \text{ } \Omega \text{ m}$
Temperature coefficient of resistance (α)	$3.9 \times 10^{-3} \text{ K}^{-1}$

- 7.6 The equivalent circuit of a pyroelectric detector may be taken to be a current source i feeding into a parallel combination of a capacitor C and a load resistor R_L (see Fig. 7.7b).

If P is the dipole moment per unit volume of the pyroelectric material, show that the surface charge is PA where A is the surface area. Hence show that i can be written

$$i = A \frac{dP}{dT} \cdot \frac{dT}{dt}$$

where T is the temperature of the detector and dT/dt the rate of increase of temperature with time due to the absorption of radiation.

By assuming the results of Problem 7.3, show that if the detector completely absorbs radiation of irradiance $W_0 + W_1 \cos(2\pi ft)$ then the voltage responsivity of the device R_v can be written

$$R_v = A \frac{dP}{dT} \frac{R_L}{G} \frac{2\pi f}{(1 + 4\pi^2 f^2 H^2 / G^2)^{1/2}} \cdot \frac{1}{(1 + 4\pi^2 f^2 R_L^2)^{1/2}}$$

Sketch the behaviour of R_v with f .

- 7.7 Estimate the minimum photon flux required for a television scanning system using a photomultiplier tube with a quantum efficiency of 0.1. The signal bandwidth is 5 MHz and a signal-to-noise ratio of 100 is required.
- 7.8 Show that when the simple bias circuit shown in Fig. 7.17 is used with a photoconductive detector to detect small signal levels, then maximum voltage output signals across R_L are obtained when R_L is equal to the detector resistance R_D .

- 7.9 Show that the gain G of a photoconductive detector can be written as $G = \tau_c / \tau_d$ where τ_c is the minority carrier lifetime, and τ_d is given by

$$\frac{1}{\tau_d} = \frac{1}{\tau_n} + \frac{1}{\tau_p}$$

τ_n and τ_p being the times taken for electrons and holes to drift across the photodetector.

- 7.10 A photoconductive detector is used in conjunction with a load resistor R_L and a detection circuit which operates over a frequency range of 0 to f_{\max} Hz. Show that the r.m.s. power fluctuation due to generation-recombination noise is given by

$$\Delta W_{g-r} = \frac{2ieR_L G}{\pi \tau_c} \tan^{-1}(2\pi f_{\max} \tau_c)$$

where i is the photoconductive current, G the photoconductive gain and τ_c the minority carrier lifetime.

Optical power of $1 \mu\text{W}$ at a wavelength of $0.9 \mu\text{m}$ falls on a photoconductive detector which is used in conjunction with a load resistor of $20 \text{ k}\Omega$ and which has a quantum efficiency of 0.7. The thickness of the photoconductive layer is $10 \mu\text{m}$ and in operation a voltage of 10 V is applied across it. Assuming the values $\tau_c = 0.5 \text{ ns}$, $\mu_e = 0.25 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ and $\mu_h = 0.05 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$, calculate the noise power expected from generation-recombination noise when the detector bandwidth is 50 MHz . Compare this with the magnitude of the shot noise power which will be present.

- 7.11 An ideal photodiode (of unit quantum efficiency) is illuminated with 10 mW of radiation at $0.8 \mu\text{m}$ wavelength; calculate the current and voltage output when the detector is used in the photoconductive and photovoltaic modes respectively. The reverse bias leakage current is 10 nA .

- 7.12 Show that the field within the depletion region in a typical p-i-n photodiode detector is such as to give rise to a saturation electron velocity (assume a linear change in field with distance across the depletion region).

- 7.13 It is desired to make a photodiode based on $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$ grown on an InP substrate. Determine a suitable composition given that it is to be used at $1.55 \mu\text{m}$. You may assume that lattice matching occurs when $y = 2.16x$, and that the lattice-matched bandgap is given by

$$E_g (\text{in eV}) = 1.36 - 0.72y + 0.12y^2$$

- 7.14 Show that the responsivity of a photodiode detector as a function of signal modulation frequency f can be written as

$$R(f) = \frac{R(0)}{(1 + 4\pi^2 f^2 C_d^2 R_L^2)^{1/2}}$$

where C_d is the diode capacitance and R_L the load resistance.

- 7.15 It is desired to make a silicon p-i-n photodiode with an area of 1 mm^2 with as fast a response time as possible when used in conjunction with a 50Ω load resistor.

Estimate the thickness of the intrinsic region required. Take $\epsilon_r = 11.8$ and v_s (the electron saturation velocity) as 10^5 m s^{-1} .

Over what wavelength range would you expect the device to be most effective?

- 7.16 When both capacitance effects and carrier transit times across the depletion region contribute to the frequency response of a photodiode, the resultant time response, τ , may be written as

$$\tau_c = \tau_c^2 + \tau_t^2$$

where τ_c and τ_t are the response times corresponding to capacitance effects and transit time effects respectively. By considering the dependence of these response times on the depletion region thickness show that the optimum width for the depletion region in a p-i-n photoconductor as far as response time is concerned is obtained when $\tau_c = \tau_t$.

REFERENCES

- 7.1 R. A. Smith, F. E. Jones and R. P. Chasmar, *The Detection and Measurement of Infrared Radiation* (2nd edn), Oxford University Press, Oxford, 1968, Section 5.9.
- 7.2 J. S. Blakemore, *Solid State Physics* (2nd edn), Saunders, Philadelphia, 1974, Section 3.3.
- 7.3 J. Pierce, 'Physical sources of noise', *Proc. IRE*, **44**, 601, 1956.
- 7.4 R. H. Kingston, *Detection of Optical and Infrared Radiation*, Springer-Verlag, Berlin, 1978, Section 5.2.
- 7.5 R. King, *Electrical Noise*, Chapman and Hall, London, 1966, Section 3.
- 7.6 R. H. Kingston, *op. cit.*, Section 6.1.
- 7.7 B. G. Streetman, *Solid State Electronic Devices* (2nd edn), Prentice Hall, Englewood Cliffs, NJ, 1980, Section 7.8.2.
- 7.8 G. E. Stillman and C. M. Wolfe, 'Avalanche photodiodes', in R. K. Willardson and A. C. Beer (eds) *Semiconductors and Semimetals*, Vol. 12, Academic Press, New York, 1977, Chapter 5.
- 7.9 B. G. Streetman, *op. cit.*, Section 9.4.3.
- 7.10 K. Chang (ed.), *Handbook of Microwave and Optical Components*, Vol. 4, *Fiber and Electro-Optical Components*, John Wiley, New York, 1991, Section 8.3.2.
- 7.11 K. Chang (ed.), *op. cit.*, Section 8.4.2.

Fiber optical waveguides

It has long been realized that light, with its carrier frequency of some 10^{14} Hz, has the potential to be modulated at much higher frequencies than either radio or microwaves and thus opens up the possibility of a single communication channel with extremely high information content. One of the main difficulties in implementing this was that without some sort of guiding medium, any form of atmospheric transmission was limited to line-of-sight communications and was thus subject to the vagaries of the weather. As long ago as 1870 Tyndall demonstrated that light could be guided within a water jet as a result of total internal reflection (ref. 8.1). However, although some theoretical studies were carried out in the early years of the present century, it was not until the mid-1960s that the idea of a communication system based on the propagation of light within circular dielectric waveguides was considered seriously (ref. 8.2). One of the reasons for this delay was that initially it was assumed the waveguides would be a dielectric rod which would carry a single electromagnetic mode. Theory showed that such guides would have an extremely small diameter and furthermore the mode fields would penetrate into the air surrounding the rod, giving rise to high losses and making it difficult to support the guide. These problems were overcome with the proposal in 1954 (ref. 8.3) to use clad dielectric waveguides. The basic structure of such a guide is shown in Fig. 8.1. It consists of a central *core* region surrounded by a *cladding*, where the core material has a higher refractive index than the cladding. To appreciate how light may travel down the waveguide we consider a ray which passes through the centre of the guide and hits the core/cladding interface at an angle that gives rise to total internal reflection (see eq. 1.14). The ray is then able to travel in a zigzag path down the core of the guide as shown in Fig. 8.1.

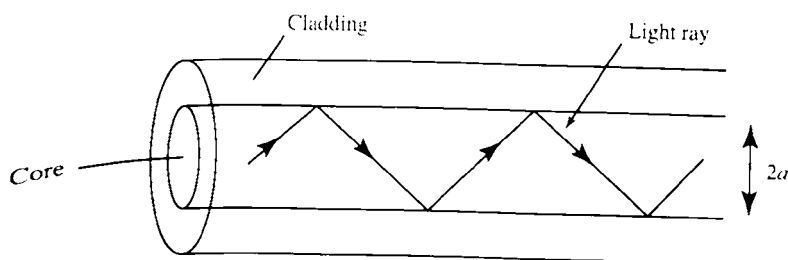


FIG. 8.1 The basic structure of an optical fiber. A core region is surrounded by a cladding region where the refractive index of the core material is greater than that of the cladding.

Although a full treatment of the propagation of light down such circular waveguides (or *optical fibers* as they are often called) requires the solution of Maxwell's equations, which is beyond the scope of the present text, it is possible to gain a reasonable understanding by using simple ray theory and to this end we take a closer look at the phenomenon of total internal reflection.

8.1 Total internal reflection

When an electromagnetic wave is incident upon the boundary between two dielectric media whose refractive indices are n_1 and n_2 , then in general a portion of that wave is reflected and the remainder transmitted. Maxwell's equations require that both the tangential components of \mathbf{E} and \mathbf{H} and the normal components of \mathbf{D} ($=\epsilon_r\epsilon_0\mathbf{E}$) and \mathbf{B} ($=\mu_r\mu_0\mathbf{H}$) are continuous across the boundary. A detailed consideration of the consequences of applying these conditions is too lengthy to be reproduced here; however, this is quite straightforward and is given in many textbooks (see e.g. ref. 8.4). The resulting equations are known as *Fresnel's equations* and are summarized below. We suppose the wave to be incident on the interface at an angle θ_i to the normal and that the reflected and transmitted waves are at angles θ_r and θ_t respectively, as shown in Fig. 8.2. These angles are related by the equations

$$\theta_i = \theta_r \quad \text{and} \quad \frac{\sin \theta_i}{\sin \theta_t} = \frac{n_2}{n_1} \quad (8.1)$$

Fresnel's equations deal with the magnitudes of the transmitted and reflected electric fields

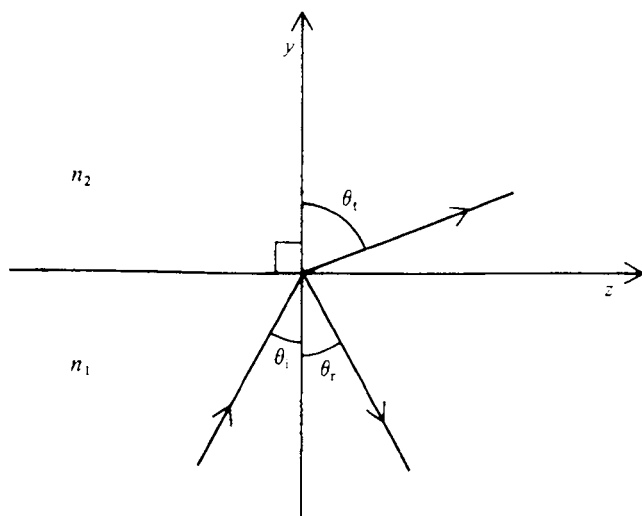


FIG. 8.2 Illustration of the behaviour of a light ray incident on the boundary between two media with refractive indices n_1 and n_2 where $n_1 > n_2$. In general, a transmitted and a reflected beam are produced. The angles of incidence, transmission and reflection are θ_i , θ_t and θ_r , respectively.

$(\mathcal{E}_r, \mathcal{E}_t)$ relative to the incident field \mathcal{E}_i . We must distinguish between the cases where the electric field vector is parallel to and perpendicular to the plane of incidence. We denote these field components by \mathcal{E}^{\parallel} and \mathcal{E}^{\perp} . It should be noted that later on we use the notations TM (Transverse Magnetic) and TE (Transverse Electric) to denote these two polarization situations. This latter notation arises because for TM polarization the magnetic field always remains perpendicular to the plane of incidence, whilst for TE polarization it is the electric field which always remains perpendicular to the plane of incidence. Fresnel's equations can then be written

$$\frac{\mathcal{E}_r^{\perp}}{\mathcal{E}_i^{\perp}} = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t}$$

$$\frac{\mathcal{E}_r^{\parallel}}{\mathcal{E}_i^{\parallel}} = \frac{n_1 \cos \theta_t - n_2 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i} \quad (8.1a)$$

and

$$\frac{\mathcal{E}_t^{\perp}}{\mathcal{E}_i^{\perp}} = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_i + n_2 \cos \theta_t}$$

$$\frac{\mathcal{E}_t^{\parallel}}{\mathcal{E}_i^{\parallel}} = \frac{2n_1 \cos \theta_t}{n_1 \cos \theta_t + n_2 \cos \theta_i} \quad (8.1b)$$

We will be particularly interested in the situation when $n_1 > n_2$, since it is then possible to have total internal reflection (see section 4.6.4); that is, when $\theta_i > \theta_c$ where

$$\theta_c = \sin^{-1}(n_2/n_1) \quad (8.2)$$

there is no transmitted wave in the second medium. Further confirmation that total internal reflection occurs may be obtained as follows: we have that

$$\cos \theta_t = (1 - \sin^2 \theta_i)^{1/2}$$

and substitution for $\sin^2 \theta_t$ from eq. (8.1) gives

$$\cos \theta_t = \left[1 - \left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i \right]^{1/2}$$

When $\theta_i > \theta_c$, $\sin \theta_i > n_2/n_1$ and $\cos \theta_t$ then becomes wholly imaginary. We may therefore write

$$\cos \theta_t = \pm iB \quad (\theta_i > \theta_c) \quad (8.3)$$

where

$$B = \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2}$$

We must take the negative sign in eq. (8.3) for reasons which will be explained later (see eq. 8.7 and the subsequent discussion).

Putting $A = (n_1/n_2) \cos \theta_i$ and substituting for A and B into the first equation of eq. (8.1a), we obtain

$$\frac{\mathcal{E}_r^\perp}{\mathcal{E}_i^\perp} = \frac{A + iB}{A - iB} \quad (8.4)$$

Inspection of the right-hand side of eq. (8.4) shows that it has a modulus of unity, which indicates that when $\theta_i > \theta_c$ the irradiances of the reflected and the incident beams are equal and thus there is no transmitted beam. In other words we have total internal reflection. Since the electric field ratios are now complex quantities, however, a *phase shift* is present between the incident and reflected beams, which we now calculate. We may rewrite eq. (8.4) as

$$\frac{\mathcal{E}_r^\perp}{\mathcal{E}_i^\perp} = \frac{\cos \psi + i \sin \psi}{\cos \psi - i \sin \psi} = \frac{\exp(i\psi)}{\exp(-i\psi)} = \exp(2i\psi) \quad (8.5)$$

where

$$\tan \psi = B/A = \frac{n_2[(n_1/n_2)^2 \sin^2 \theta_i - 1]^{1/2}}{n_1 \cos \theta_i}$$

or

$$\tan \psi = \frac{[\sin^2 \theta_i - (n_2/n_1)^2]^{1/2}}{\cos \theta_i} \quad (8.5a)$$

Similarly the second equation of eq. (8.1a) yields

$$\frac{\mathcal{E}_r^\parallel}{\mathcal{E}_i^\parallel} = \exp(2i\delta) \quad (8.6)$$

where

$$\tan \delta = (n_1/n_2)^2 \tan \psi \quad (8.6a)$$

The phase changes on reflection for \mathcal{E}^\perp and \mathcal{E}^\parallel are thus given by 2ψ and 2δ respectively; in both cases \mathcal{E}_r leads \mathcal{E}_i in phase. The variation of both these phase changes as a function of θ_i is shown for a glass/air interface (i.e. $n_1 = 1.5$, $n_2 = 1$) in Fig. 8.3.

Although all the energy in the beam is reflected when $\theta_i > \theta_c$ there is still a disturbance in the second medium whose electric field amplitude decays exponentially with distance away from the boundary. No energy is conveyed away from the surface provided the second medium extends an infinite distance from the boundary. We may derive an expression for this decay by considering the phase factor \mathcal{P} of the transmitted wave, which at a point \mathbf{r} we may write as

$$\mathcal{P} = \exp[i(\omega t - \mathbf{k}_t \cdot \mathbf{r})]$$

where \mathbf{k}_t is the wavevector associated with the transmitted wave. Reference to Fig. 8.4 shows

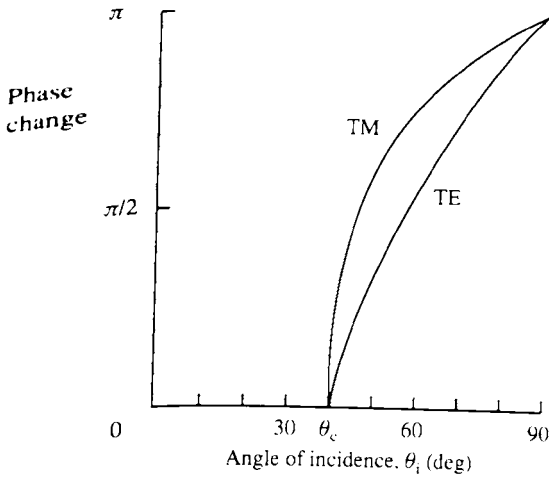


FIG. 8.3 The variation of the phase shift on reflection as a function of the angle of incidence at an interface where there is total internal reflection.

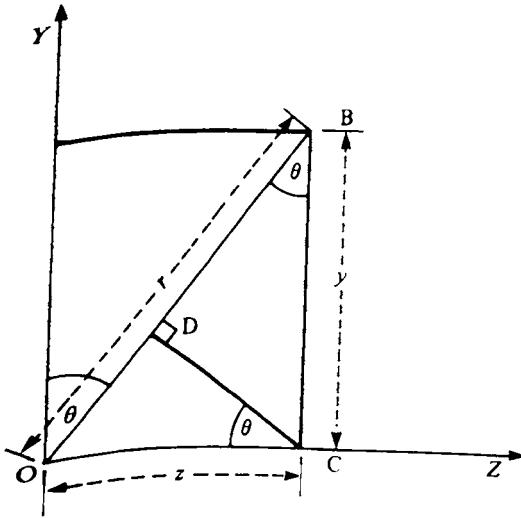


FIG. 8.4 Illustration of the relationship between the rectangular coordinates y and z and the distance r measured from the origin O . We have $OB = OD + DB = OC \sin \theta + BC \cos \theta$. Hence, in terms of the coordinates, $r = z \sin \theta + y \cos \theta$.

that we may write r as $z \sin \theta_i + y \cos \theta_i$, and hence we have

$$\begin{aligned} \mathcal{P} &= \exp\{i[\omega t - k_1(z \sin \theta_i + y \cos \theta_i)]\} \\ &= \exp\left[i\left(\omega t - \frac{2\pi n_2}{\lambda_0}(z \sin \theta_i + y \cos \theta_i)\right)\right] \end{aligned}$$

where λ_0 is the wavelength of the radiation in vacuum. If we substitute expressions for $\sin \theta_i$ and $\cos \theta_i$ from eqs (8.1) and (8.3), we obtain

$$\begin{aligned} \psi &= \exp \left\{ i \left[\omega t - \frac{2\pi n_2}{\lambda_0} \left(z \frac{n_1}{n_2} \sin \theta_i + (\pm iB)y \right) \right] \right\} \\ &= \exp \left(\pm B \frac{2\pi n_2}{\lambda_0} y \right) \exp \left[i \left(\omega t - \frac{2\pi n_1 \sin \theta_i}{\lambda_0} z \right) \right] \end{aligned} \quad (8.7)$$

Thus, in the y direction the wave either grows or decays exponentially with distance. The former situation is obviously a non-physical solution and we must choose $\cos \theta_i = -iB$ in eq. (8.3).

The decay with distance in the second medium is given by the factor $F(y)$ where

$$\begin{aligned} F(y) &= \exp \left(- \frac{2\pi n_2}{\lambda_0} B y \right) \\ &= \exp \left\{ - \frac{2\pi n_2}{\lambda_0} \left[\left(\frac{n_1}{n_2} \right)^2 \sin^2 \theta_i - 1 \right]^{1/2} y \right\} \end{aligned} \quad (8.8)$$

Usually, $F(y)$ decays rapidly with y (see Example 8.1). However, when θ_i is very close to θ_c then $[(n_1/n_2)^2 \sin^2 \theta_i - 1]^{1/2}$ will be close to zero and the disturbance may extend an appreciable distance into the second medium. That part of the field which is in the second medium is referred to as the *evanescent field*.

EXAMPLE 8.1 Field penetration into the less dense medium during total internal reflection —

If we again take a glass/air interface, $n_1 = 1.5$, $n_2 = 1$, with $\theta_i = 60^\circ$, from eq. (8.8) we obtain

$$F(y) = \exp(-5.1y/\lambda_0)$$

Thus, in a distance equal to the wavelength, the magnitude of the electric field will fall by a factor $\exp(-5.1)$ or 5.5×10^{-3} . If, however, we take a value for θ_i which is much closer to $\theta_c (= 41.8^\circ)$, say 42° , then we have

$$F(y) = \exp(-0.54y/\lambda_0)$$

As a function of distance the decay is now less rapid than before, and in fact the field decay factor at a distance $y = \lambda_0$ is now $\exp(-0.54)$ or 0.58.

8.2

Planar dielectric waveguide

At this stage it is useful to examine one of the simplest forms of waveguide, namely the planar symmetric dielectric waveguide. There are two main reasons for this. First it is possible to make considerable progress in understanding the behaviour of these guides using ray theory.

and secondly many of the resulting ideas can be applied, with relatively minor modifications, to circular waveguides. In addition planar waveguides are of practical importance in their own right: for example, they form the basis of the waveguides used in integrated optoelectronics (see section 9.4). The waveguide itself consists of a semi-infinite slab of dielectric of thickness d and refractive index n_1 sandwiched between two regions both of refractive index n_2 , where $n_1 > n_2$ (Fig. 8.5). The central slab is called the *core* whilst the surrounding regions are called the *cladding* regions. A ray of light may readily propagate down the core by taking the zigzag path as shown in Fig. 8.6, provided that total internal reflection occurs at the core/cladding interfaces. Thus we require that $90^\circ > \theta > \theta_c$, where θ is the internal ray angle (note that in the interests of notational simplicity, from now on we refer to the internal ray angle as θ rather than θ_i).

Although in Fig. 8.6 we have drawn only one ray, we must realize that in fact an 'infinite' number of such rays, all slightly displaced from each other, will also be propagating down

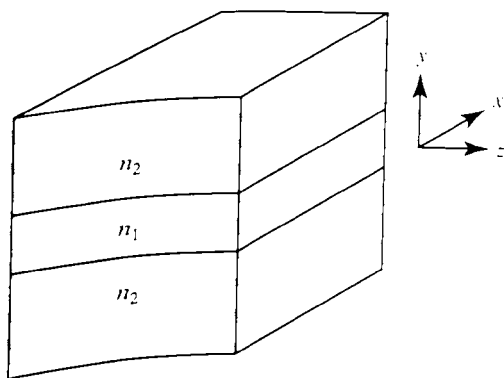


FIG. 8.5 Structure of a planar symmetric dielectric waveguide. A layer of refractive index n_1 is sandwiched between layers of refractive index n_2 where $n_1 > n_2$.

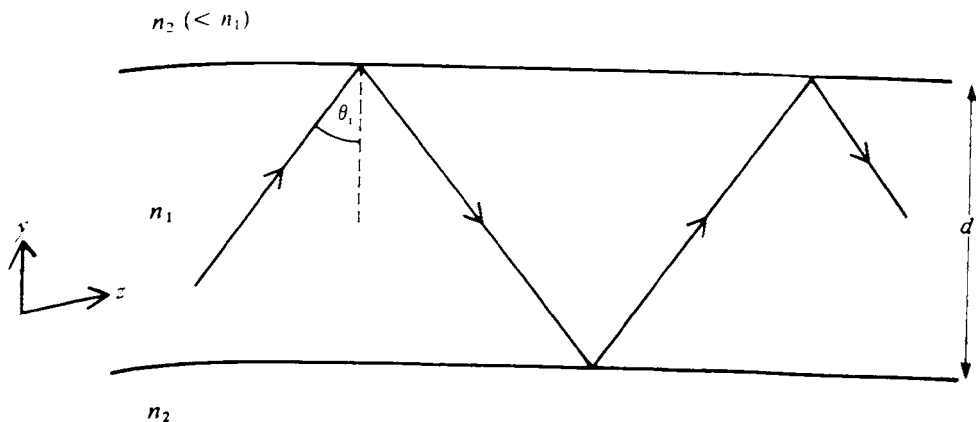


FIG. 8.6 The zigzag path of a light ray down a planar dielectric waveguide that results when the angle of incidence at the boundary θ_i is greater than the critical angle θ_c .

the guide. Actually, the rays merely represent lines drawn normally to the plane wavefronts. Figure 8.7(a) shows a wavefront for the situation where the ray angle, θ , is less than 45° whilst Fig. 8.7(b) shows the situation where θ is greater than 45° . We deal with the former situation first. All points along the same wavefront must have identical phases and if we consider the particular wavefront, FC, drawn in Fig. 8.7(a), we see that it intersects two of the upwardly travelling portions of the same ray at the points A and C. Unless the phase at the points C and A is the same, or differs by a multiple of 2π , destructive interference will take place, making it impossible for light to propagate down the guide. To calculate the phase difference between A and C we must take into account two factors, first the path length of $AB + BC$ and secondly the phase changes due to reflection at B and C. For convenience we may write the phase change on reflection simply as $\phi(\theta)$ where, for TE radiation, $\phi(\theta) = 2\psi$ and, for TM radiation, $\phi(\theta) = 2\delta$ (see eqs 8.5 and 8.6). Thus the total phase change may be written

$$(AB + BC) \frac{2\pi n_1}{\lambda_{c1}} - 2\phi(\theta)$$

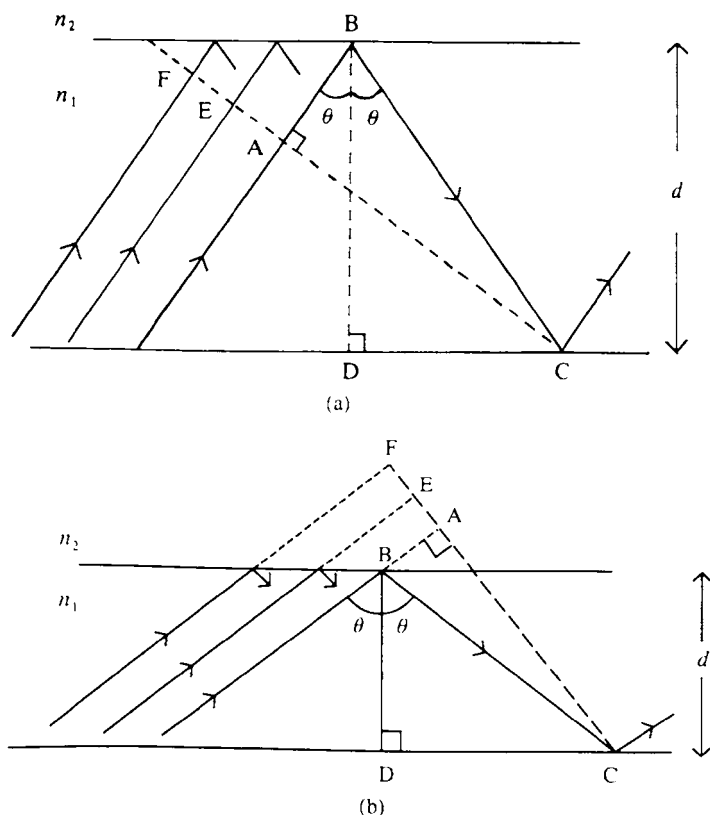


FIG. 8.7 Ray paths having the same internal angle θ within a planar waveguide. A wavefront is shown connecting the points F, E, A, C, which must therefore have the same phase.

Note that the two phase terms have opposite sign. This is because the basic equation of a wave $A = A_0 \cos(\omega t - kr)$ implies that the phase at the point r_2 will be less than that at a point r_1 if $r_2 > r_1$, that is the phase *decreases* with increasing r , whereas when a wave suffers total internal reflection its phase *increases* (see eqs 8.5 and 8.6).

From the triangle ABC we have that $AB = BC \cos 2\theta$, and hence $AB + BC = BC(1 + \cos 2\theta)$ or, since $\cos 2\theta = 2 \cos^2 \theta - 1$, $AB + BC = 2BC \cos^2 \theta$. Also, from the triangle BDC we have $BC \cos \theta = d$, so that finally, $AB + BC = 2d \cos \theta$.

The condition for the mode to propagate is therefore

$$\frac{4\pi n_1 d \cos \theta}{\lambda_0} - 2\phi(\theta) = 2m\pi$$

or

$$\frac{2\pi n_1 d \cos \theta}{\lambda_0} - \phi(\theta) = m\pi \quad (8.9)$$

where m is an integer.

So far the discussion has only covered the situation where the internal ray angle in the guide is less than 45° . Turning to Fig. 8.7(b), we again require that the phases at A and C be equal, but this time the distance AB represents a *negative* path length since the ray would have to travel 'backwards' along AB to get from A to B. Thus the total phase change between A and C can be written

$$(BC - AB) \frac{2\pi n_1}{\lambda_0} - 2\phi(\theta)$$

It is left as an exercise for the reader to show that in Fig. 8.7(b) $(BC - AB)$ is equal to $2d \cos \theta$, so that in fact the phase condition represented by eq. (8.9) applies for all values of the internal ray angle θ .

For each value of m there will be a corresponding value of θ , namely θ_m , that satisfies eq. (8.9). The difficulty is that the dependence of $\phi(\theta)$ on θ (eqs 8.5a and 8.6a) is such that we cannot obtain an explicit expression for θ in terms of m . Equation (8.9) may, however, be solved either graphically or numerically and it is useful to consider a graphical solution here. Figure 8.8 shows graphs of the equations

$$y = \pi m + \phi(\theta) \quad (8.10)$$

for several different values of m . Also plotted is the curve

$$y = \frac{2\pi d n_1 \cos \theta}{\lambda_0} \quad (8.11)$$

Evidently the intersection of the two curves given by eqs (8.10) and (8.11) results in a value of θ which will solve eq. (8.9). Note, however, that we are only interested in solutions where the value of θ is greater than the critical angle θ_c , otherwise the ray will not be totally internally reflected at the interface.

Thus for each value of m there is at most one value of θ (i.e. θ_m) which will solve eq. (8.9).

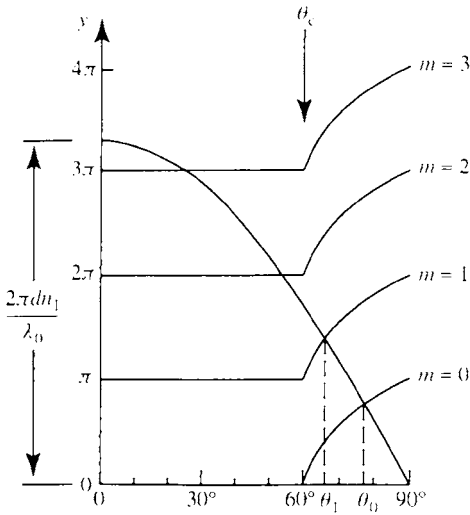


FIG. 8.8 Illustration of a graphical solution to eq. (8.9). For the particular waveguide dimensions and light wavelength used here there are only two possible ray angles which satisfy both the phase requirements of eq. (8.9) and the condition for total internal reflection: these are designated by θ_0 and θ_1 .

Each value of θ_m ($> \theta_c$) is, as we shall see later, associated with a distinct distribution of electric field across the guide. Such a distribution is referred to as a *mode*. (We have met the idea of modes previously when dealing with lasers in Chapter 5. Indeed the present discussion is directly relevant to the possible modes in a semiconductor laser.) The characteristics of the mode depend not only on the value of m but also on whether the phase shifts on reflection result from the \mathcal{E}^{\parallel} and \mathcal{E}^{\perp} situations. Since these two polarizations are also referred to by the symbols TM and TE, the resulting modes are known as TE_m and TM_m modes. Thus a TM_m mode involves the field components \mathcal{H}_x , \mathcal{E}_y and \mathcal{E}_z , whilst a TE_m mode involves the field components \mathcal{E}_x , \mathcal{H}_y and \mathcal{H}_z (Fig. 8.9).

If a particular mode has a value of θ_m where $\theta_m = \theta_c$, we say that the mode is *at cut-off*. If $\theta_m < \theta_c$ the mode is *below cut-off* (and in consequence will be rapidly attenuated and hence not propagate for any appreciable distance), whilst if $\theta_m > \theta_c$ the mode is *above cut-off*. It is evident from Fig. 8.8 that both TE_m and TM_m modes will be at cut-off if

$$\frac{2\pi dn_1 \cos \theta_c}{\lambda_0} = \pi m \quad (8.12)$$

Since $\cos \theta_c = (1 - \sin^2 \theta_c)^{1/2} = [1 - (n_2/n_1)^2]^{1/2}$ we will have mode cut-off when

$$\frac{2\pi dn_1}{\lambda_0} \left[1 - \left(\frac{n_2}{n_1} \right)^2 \right]^{1/2} = \pi m$$

or

$$V = \frac{\pi}{2} m \quad (8.13)$$

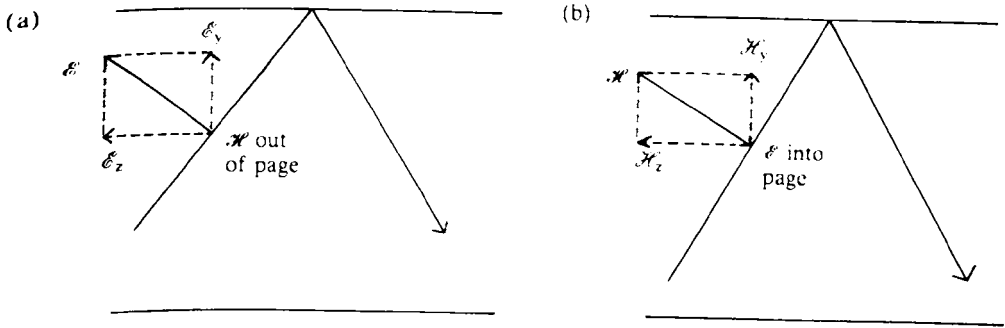


FIG. 8.9 A ray with \mathcal{E} in the plane of incidence (a) gives rise to the three field components \mathcal{H}_y , \mathcal{E}_x and \mathcal{E}_z corresponding to a TM mode. A ray with \mathcal{H} in the plane of incidence (b) gives rise to \mathcal{E}_y , \mathcal{H}_x and \mathcal{H}_z and corresponds to a TE mode.

where

$$V = \frac{\pi d n_1}{\lambda_0} \left[1 - \left(\frac{n_2}{n_1} \right)^2 \right]^{1/2} = \frac{\pi d}{\lambda_0} (n_1^2 - n_2^2)^{1/2} \quad (8.14)$$

The parameter V is referred to in a variety of ways:¹ simply as the V parameter, the *normalized film thickness*, or the *normalized frequency*.

It is evident then that the number of guided TE, or TM, modes in a guide, N , is given by

$$N = 1 + \text{INT}(2V/\pi) \quad (8.15)$$

where INT indicates that the integer part of the following expression is taken.

EXAMPLE 8.2 Number of modes in a planar dielectric guide

We consider a planar dielectric waveguide $100 \mu\text{m}$ thick where $n_1 = 1.48$, $n_2 = 1.46$ and where light of wavelength $1 \mu\text{m}$ is being used. For this guide

$$V = \frac{\pi \times 100 \times 10^{-6}}{1 \times 10^{-6}} (1.48^2 - 1.46^2)^{1/2} = 76.2$$

Thus $2V/\pi = 48.5$. Equation (8.15) then gives the total number of possible modes (both TE and TM) as $2 \times [1 + \text{INT}(48.5)] = 2 \times (1 + 48) = 98$.

An interesting result that follows directly from eq. (8.15) is that only one mode² (i.e. the $m = 0$ mode) will propagate if

$$V < \pi/2 \quad (8.16)$$

Such a guide is called a *single mode guide*. This result implies that the $m = 0$ mode has no cut-off, so that, in theory, whatever the dimensions of the guide the lowest order mode will always propagate. (This is not the case if the guide is asymmetric; that is, it has media of differing refractive indices above and below the central layer, see section 9.4.1.)

EXAMPLE 8.3 Single mode guide dimensions

From eq. (8.16) the condition for single mode behaviour can be written as

$$\frac{d}{\lambda_0} < \frac{1}{2(n_1^2 - n_2^2)^{1/2}}$$

Again taking a guide where $n_1 = 1.48$ and $n_2 = 1.46$ and using a wavelength of $1\text{ }\mu\text{m}$, the guide has to have a thickness d such that

$$d < \frac{1}{2(1.48^2 - 1.46^2)^{1/2}} \mu\text{m}$$

That is, the waveguide core thickness must be less than $2.06\text{ }\mu\text{m}$.

As we have seen, a given mode can be regarded as being made up from an 'infinite' collection of rays travelling down the guide with the same value of internal angle θ . If we consider any particular point within the guide, only two of these rays can pass through it. One will be directed 'upwards', the other directed 'downwards'. Since, in general, there will be a phase difference between these two, they will interfere and thereby give rise to a variation in the field amplitude across the guide. Each mode is characterized by a different field variation.

This phase difference, as a function of position within the guide, may be determined from Fig. 8.10. This shows two rays meeting at a point C, a distance y above the centre of the guide. The line AC represents a wavefront, and hence the phases at the points A and C must be the same. Thus the phase difference between the two rays meeting at C, $\Delta\Phi(y)$, arises from a path difference of $AB + BC$ together with a phase change of $\phi(\theta)$ on reflection at B.³

We have that $AB = BC \cos 2\theta$, so that $AB + BC = 2BC \cos^2 \theta$ (as shown above). Since we

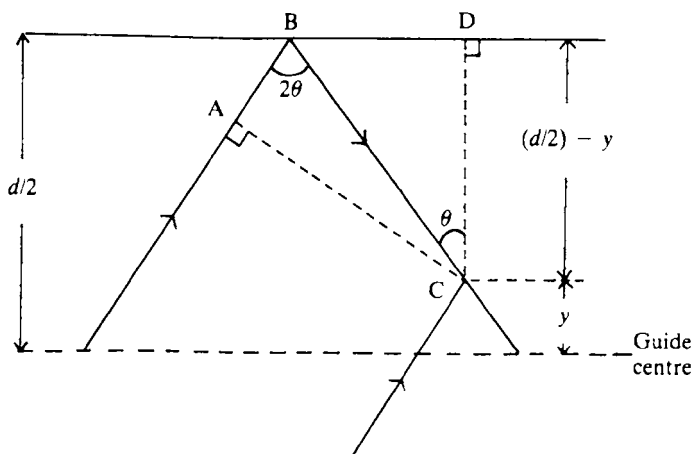


FIG. 8.10 Two 'oppositely' directed rays meeting at a point C a distance y above the guide centre.

also have $BC = [(d/2) - y] \cos \theta$, it follows that $AB + BC = 2[(d/2) - y] \cos \theta$. Thus we may write

$$\Delta\Phi(y) = 2\left(\frac{d}{2} - y\right) \frac{2\pi n_1 \cos \theta_m}{\lambda_0} - \phi(\theta)$$

By substituting for $\cos \theta_m$ from eq. (8.9) we obtain

$$\Delta\Phi(y) = 2\left(\frac{d}{2} - y\right) \frac{2\pi n_1}{\lambda_0} \frac{[m\pi + \phi(\theta)]\lambda_0}{2\pi d n_1}$$

or

$$\Delta\Phi(y) = m\pi - \frac{2y}{d} [m\pi + \phi(\theta)] \quad (8.17)$$

Now the resultant of two waves with a phase difference $\Delta\Phi(y)$ may be written as

$$\mathcal{E}_0 \{ \cos(\omega t) + \cos[\omega t + \Delta\Phi(y)] \}$$

or

$$2\mathcal{E}_0 \cos\left(\omega t + \frac{\Delta\Phi(y)}{2}\right) \cos\left(\frac{\Delta\Phi(y)}{2}\right)$$

The effective amplitude of the electric field is thus given by $2\mathcal{E}_0 \cos[\Delta\Phi(y)/2]$ which from eq. (8.17) can be written as

$$2\mathcal{E}_0 \cos\left(\frac{m\pi}{2} - \frac{y}{d} [m\pi + \phi(\theta)]\right) \quad (8.18)$$

At the centre of the guide ($y=0$) this will have the value $\pm 2\mathcal{E}_0$ or zero, depending on whether m is even or odd. At the guide edges ($y=d/2$) the amplitude is $2\mathcal{E}_0 \cos[\phi(\theta)]$. The complete mode field pattern is obtained by matching this sinusoidal variation in the core with the exponential decline in the cladding as given by eq. (8.8). Figure 8.11 shows the variation in electric field amplitude across both the core and the cladding for the first four TE modes in a planar dielectric waveguide. In fact for the particular waveguide parameters chosen, these are the *only* TE modes allowed to propagate.

Figure 8.12 illustrates the mode field of the TE_1 mode both close to and far from cut-off. It is seen that a large proportion of the mode field extends into the cladding near cut-off. In fact cut-off marks the point when all the mode energy resides in the cladding. This behaviour is readily understood from an inspection of Fig. 8.8 since, as V decreases and cut-off is approached, θ_m will get closer to θ_c and hence both ψ and δ (and hence $\phi(\theta)$) will tend towards zero. This in turn will cause the field amplitude at the cladding boundary to increase. In addition we have already noted that the rapidity of decay of the evanescent field increases as the guide angle θ approaches θ_c (Example 8.1). A similar trend can be observed in Fig. 8.11: as the mode number increases the proportion of the mode field that is in the cladding increases. This again may be understood by noting that as m increases θ_m will get closer to θ_c .

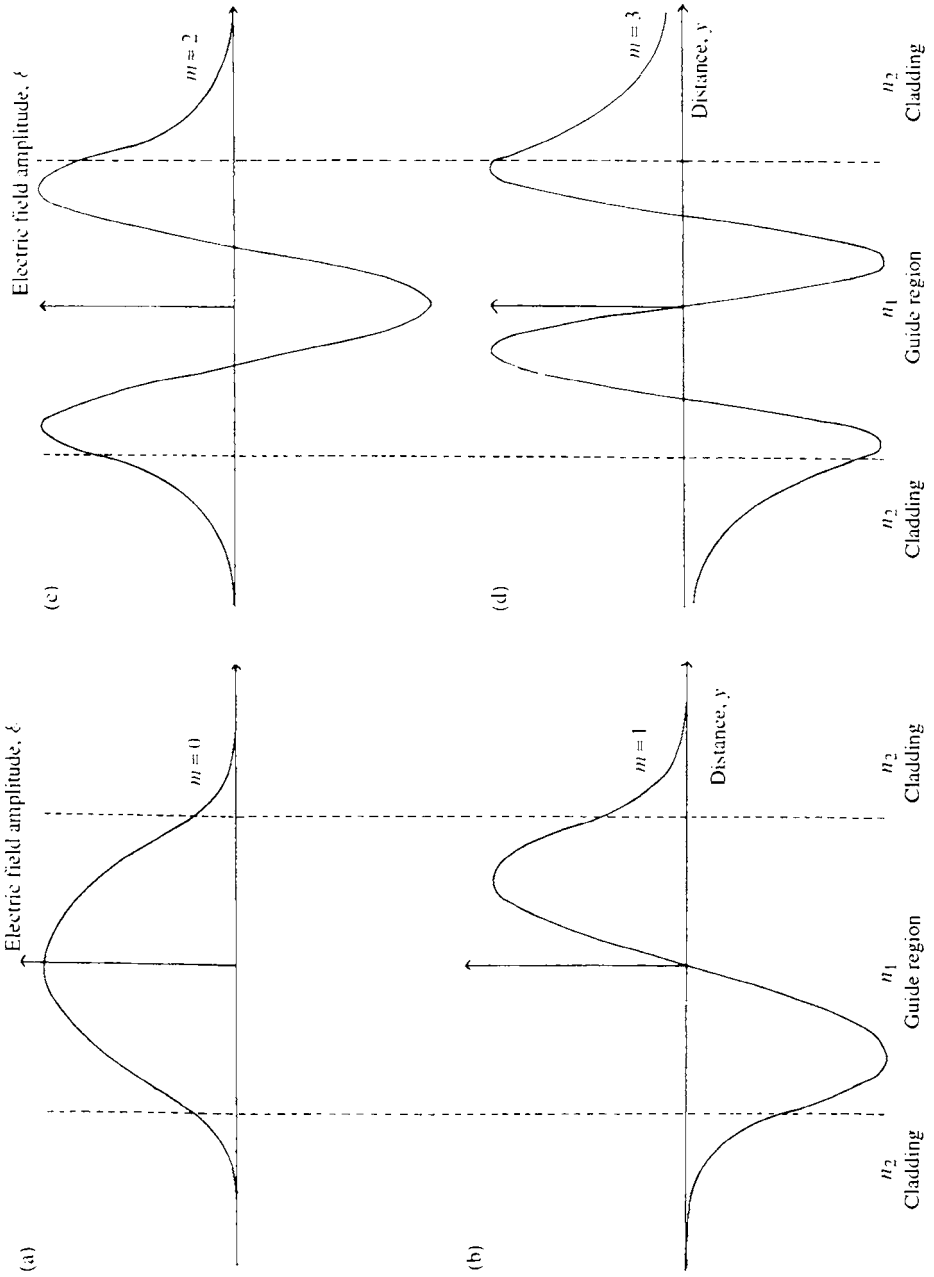


FIG. 8.11 Transverse electric field amplitudes for the four lowest modes in a symmetric planar waveguide, where $n_1 = 1.5$, $n_2 = 1$, $d = 2 \mu\text{m}$ and $\lambda_0 = 1.2 \mu\text{m}$.

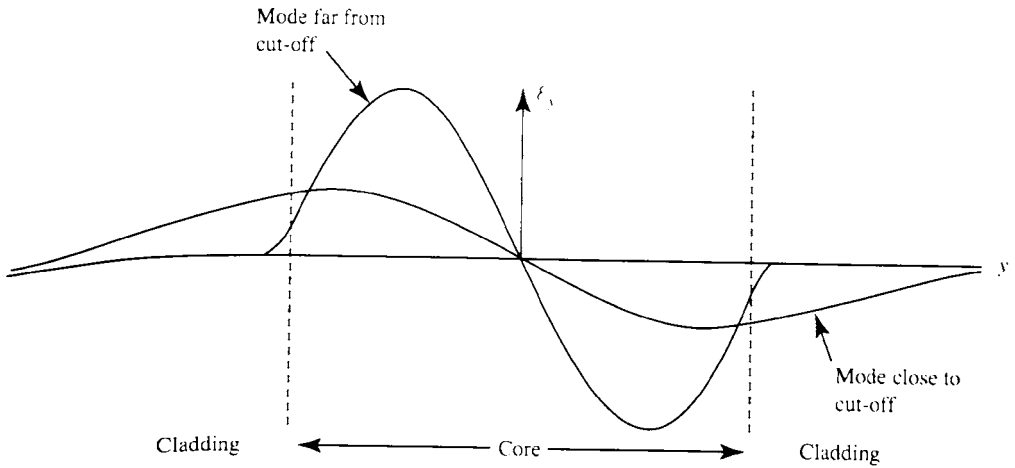


FIG. 8.12 The mode field of TE_1 when it is close to cut-off and far from cut-off. When close to cut-off the mode field spreads out an appreciable distance into the cladding.

8.3

Optical fiber waveguides

In their simplest form optical fibers have a cylindrical geometry with a core region surrounded by a cladding (Fig. 8.1). As in the planar waveguide the refractive index of the core is greater than that of the cladding. Such fibers are referred to as *step index* fibers because of the shape of the refractive index profile (Fig. 8.13). The resulting structure is often covered by an additional coating layer (usually of plastic) which serves to strengthen the fiber and to provide protection against chemical attack. As with the planar waveguide, radiation is restricted to travel down the guide in certain 'modes', with each mode corresponding to a distinct electromagnetic field distribution across the guide. The equations describing the modes in circular waveguides are much more complicated than their planar waveguide counterparts, and

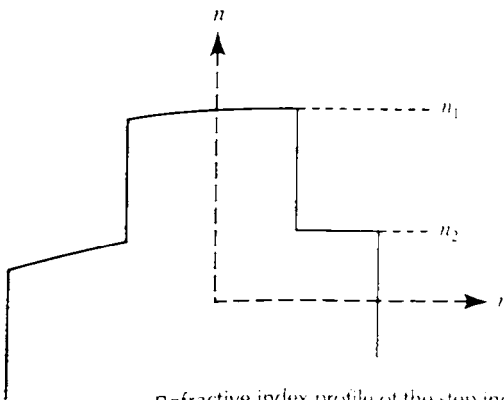


FIG. 8.13 Refractive index profile of the step index fiber.

we make no attempt to derive them here (see e.g. ref. 8.5). However, we can take a number of the concepts developed for planar waveguides and apply them, with some modification, to circular waveguides.

We may divide the rays travelling within the fiber into two types, *meridional* rays and *skew* rays as illustrated in Figs 8.14(a) and 8.14(b) respectively. The former pass through the central axis of the guide and are in many ways similar to the rays in the planar waveguide. As in the planar waveguide they give rise to modes which are designated transverse electric (TE) and transverse magnetic (TM). In this case, however, two integers l and m are required to specify completely the mode rather than just one (m) as before (essentially because of the two-dimensional nature of the guide). We thus refer to TE_{lm} and TM_{lm} modes.

Skew rays have no analog in planar waveguides and describe angular 'helices' within the guide. Because of the angles involved, components of *both* \mathcal{E} and \mathcal{H} can be transverse to the fiber axis. Consequently the modes originating from skew rays are designated as either HE_{lm} or EH_{lm} depending on whether their magnetic or electric characters are more important. It is also possible to launch some skew rays which, strictly speaking, do not correspond to bound modes, but which nevertheless can propagate for appreciable distances before being lost from the fiber. The resulting modes are called *leaky* modes.

In most practical waveguides, the refractive indices of the core and cladding differ from each other by only a few per cent and it may be shown (ref. 8.6) that the full set of modes (i.e. TE_{lm} , TM_{lm} , HE_{lm} and EH_{lm}) can be approximated by a set of so-called *linearly polarized* (LP_{lm}) modes. The electric field intensity profiles of three such modes are illustrated in Fig. 8.15. An LP_{lm} mode in general has m field maxima along a radius vector and $2l$ field maxima round a circumference. On a ray picture, l is a measure of the degree of helical propagation: the larger its value the tighter the helix. The integer m , on the other hand, is related to the angle θ : the larger m the smaller the value of θ involved.

As in planar waveguides (see eq. 8.14) the number of modes in a step index circular

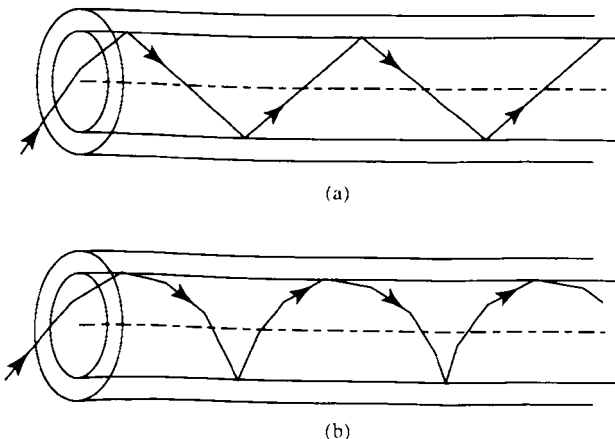


FIG. 8.14 Ray paths in step index fibers: (a) meridional rays; (b) skew rays.

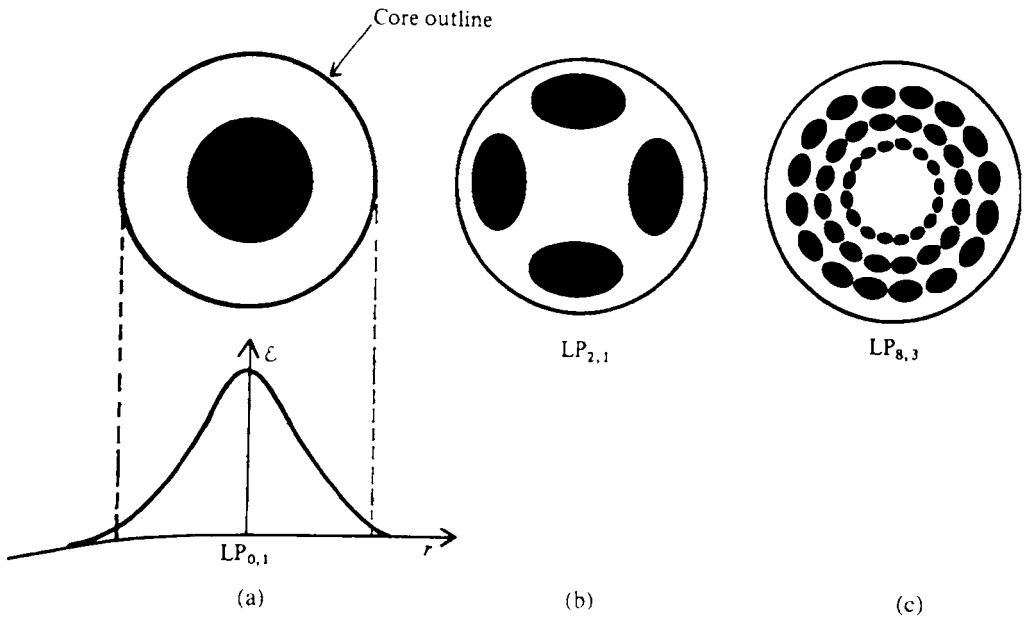


FIG. 8.15 Electric field distributions in three circular waveguide modes.

waveguide is determined by a V parameter, where now

$$V = \frac{2\pi a}{\lambda_0} (n_1^2 - n_2^2)^{1/2} \quad (8.19)$$

The fiber can support only one mode when $V < 2.405$ whilst when $V \gg 1$ the number of modes, N , that can propagate is given by

$$N \approx V^2/2 \quad (8.20)$$

Apart from very special circumstances fibers are either *multimode*, where the fiber core diameter is such that the number of modes is very large (e.g. see Example 8.4), or *single mode* when only one mode can propagate. In multimode fibers the modes are so numerous that they effectively involve a continuum of ray angles within the guide. This enables an analysis of some of the guide properties to be carried out using ray theory.

EXAMPLE 8.4 Number of modes in a fiber

We consider a fiber where $n_1 = 1.48$ and $n_2 = 1.46$ and $\lambda_0 = 900$ nm. If the core radius is 100 μm , then the value of V is given by

$$V = \frac{2\pi \times 100 \times 10^{-6} (1.48^2 - 1.46^2)^{1/2}}{(0.9 \times 10^{-6})} = 169.3$$

Since $V \gg 1$, the number of modes able to propagate is given, approximately, by $N \approx (169.3)^2/2 = 14\,329$.

8.3.1 Step index multimode fibers

We assume in our analysis that we need to deal only with meridional rays. This approximation owes more than a little to the fact that skew rays involve some rather unpleasant geometry! However, in practice the approximation yields reasonably acceptable results. The path of a meridional ray entering a fiber and undergoing total internal reflection is illustrated in Fig. 8.16. The angle α that the ray in the external medium (usually air) makes with the normal to the end of the fiber is related to the internal angle θ by Snell's law, so that

$$\frac{\sin \alpha}{\sin(90^\circ - \theta)} = \frac{n_1}{n_0}$$

Hence

$$\sin \alpha = \frac{n_1}{n_0} \cos \theta$$

The maximum value that α can take, α_{\max} , is thus determined by the minimum value that θ can take, which is of course the critical angle θ_c . Thus we have

$$n_0 \sin \alpha_{\max} = n_1(1 - \sin^2 \theta_c)^{1/2} = (n_1^2 - n_2^2)^{1/2}$$

The quantity $(n_1^2 - n_2^2)^{1/2}$ is known as the *numerical aperture* (NA) of the fiber, and hence

$$\alpha_{\max} = \sin^{-1}(\text{NA}/n_0) \quad (8.21)$$

α_{\max} is known as the *fiber acceptance angle*. Another parameter that is used in connection with step index fibers is Δ , where

$$\Delta = (n_1^2 - n_2^2)/2n_1^2 \quad (8.22)$$

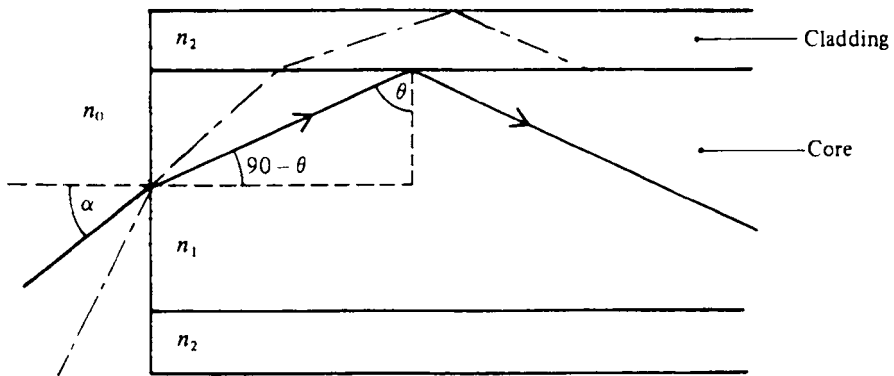


FIG. 8.16 Illustration of the path of a meridional ray (solid line) as it enters a circular step index waveguide from a medium of refractive index n_0 . The ray is incident on the end of the fiber at an angle α to the normal. Inside the waveguide, the ray makes an angle θ with the normal to the guide axis. Also shown (chain line) is the path of a ray that does not undergo total internal reflection at the core/cladding interface and does not, therefore, correspond to a bound mode.

Since n_1 and n_2 usually only differ by a few per cent, we may write

$$\Delta \approx (n_1 - n_2)/n_1 \quad (n_1 \approx n_2) \quad (8.22a)$$

The numerical aperture can be expressed in terms of Δ by

$$\text{NA} = n_1(2\Delta)^{1/2} \quad (8.23)$$

It should be noted that skew rays can have larger launch angles than meridional rays, so that, strictly speaking, the maximum acceptance angle of a fiber will be somewhat larger than that given by eq. (8.21). For most practical purposes this extra complication is usually ignored.

Figure 8.16 also shows a ray which is incident on the fiber end at an angle greater than α_{\max} . This ray is refracted into the cladding and does not correspond to a bound mode. It may, however, undergo total internal reflection at the cladding/coating (or cladding/air) interface and thus be able to propagate down the fiber to some extent, although the attenuation is expected to be high. This is an example of a *cladding mode* and, although lossy, such modes can give rise to misleading results when short lengths of fiber are used to determine such parameters as fiber attenuation and numerical aperture (see section 8.6).

EXAMPLE 8.5 Calculation of typical fiber parameters

We consider a multimode fiber in air with $n_1 = 1.48$, $n_2 = 1.46$ and $n_0 = 1$. The smallest value of the internal angle is θ_c , which is given by $\sin \theta_c = 1.46/1.48$, whence

$$\theta_c = 80.69^\circ$$

Using eq. (8.22) the parameter Δ is given by

$$\Delta = (1.48^2 - 1.46^2)/(2 \times 1.48^2) = 0.01342$$

If the approximate expression, eq. (8.22a), is used instead we have

$$\Delta \approx (1.48 - 1.46)/1.48 = 0.01351$$

We also have

$$\text{NA} = (1.48^2 - 1.46^2)^{1/2} = 0.242$$

The value of α_{\max} is then given by eq. (8.21) to be

$$\alpha_{\max} = \sin^{-1}(0.242) = 14.0^\circ$$

8.3.2 Intermodal dispersion

We know that on a simple ray model the rays corresponding to TE and TM modes are meridional rays which follow a simple zigzag path down the fiber core, very similar in fact to the ray paths in the planar waveguide (Fig. 8.6). It is evident from this diagram that the speed with which the ray travels down the waveguide depends on its internal angle. Thus along

the ray path the velocity will have a magnitude of c/n_1 , whilst the component of this velocity along the fiber axis is $(c/n_1)\cos(90^\circ - \theta)$, or $(c/n_1)\sin \theta$. The highest component velocity will be given when $\theta = 90^\circ$ whilst the smallest will occur when $\theta = \theta_c$. If we consider a length L of fiber, then the ray with the highest component velocity will take a time $\Delta\tau_{\min}$ to traverse the fiber where

$$\Delta\tau_{\min} = \frac{Ln_1}{c \sin 90^\circ} = \frac{Ln_1}{c}$$

Similarly the ray with the lowest component velocity will take a time $\Delta\tau_{\max}$ where

$$\Delta\tau_{\max} = \frac{Ln_1}{c \sin \theta_c} = \frac{Ln_1^2}{cn_2}$$

The difference between these two may be written

$$\begin{aligned} \Delta\tau^{\text{SI}} &= \Delta\tau_{\max} - \Delta\tau_{\min} \\ \Delta\tau^{\text{SI}} &= \frac{L}{c} \frac{n_1}{n_2} (n_1 - n_2) \end{aligned} \quad (8.24)$$

Thus if a pulse with an extremely narrow temporal width were to be sent down the fiber we would expect it to emerge from the other end with a maximum pulse width of $\Delta\tau^{\text{SI}}$ where $\Delta\tau^{\text{SI}}$ is given by eq. (8.24). Since we associate different ray angles with different modes, an alternative way of discussing the broadening is to say that different modes travel down the fiber with different velocities. This helps to explain the name given to the phenomenon of *intermodal dispersion*.

It is worthwhile pursuing the mode velocity concept a little further. If distance along the fiber is measured by z then the mode fields travel down the fiber with a z and t dependence which can be written as $\exp[i(\omega t - \beta z)]$, where β is the mode propagation constant (for plane waves the corresponding parameter is usually written as k , see eq. 1.6, and is known as the wavenumber). The mode phase velocity is then ω/β , but the velocity with which energy can travel down the fiber is given by the group velocity $\partial\omega/\partial\beta$ (see eq. 1.8). A schematic variation of ω with β for a particular TE or TM mode is shown in Fig. 8.17(a). At large values of ω (or β) far from mode cut-off, the group velocity tends towards the value c/n_1 . As we move down the curve towards cut-off, the group velocity slowly decreases at first, to be followed by a final rapid increase to the value c/n_2 , as cut-off is approached (Fig. 8.17b).

We can gain a qualitative understanding of most of this behaviour by using the simple ray picture. Far from cut-off, a mode has an internal guide angle that approaches 90° , and hence its effective velocity down the guide should indeed approach c/n_1 . As we move towards cut-off, the ray angle decreases and, because of the increased path lengths involved, the effective velocity down the guide decreases. However, the final rapid increase in velocity for the lowest possible β values cannot be explained using this simple picture. The discrepancy arises essentially from our neglect of the phase change on reflection. This may be incorporated into the ray model by assuming that the ray penetrates some way into the cladding at each 'reflection' as shown in Fig. 8.18. Close to cut-off the ray spends an increasing amount of time in the cladding where it travels faster than it would in the core, thus giving rise to an increase in the overall velocity down the guide. It should be mentioned that some EH and

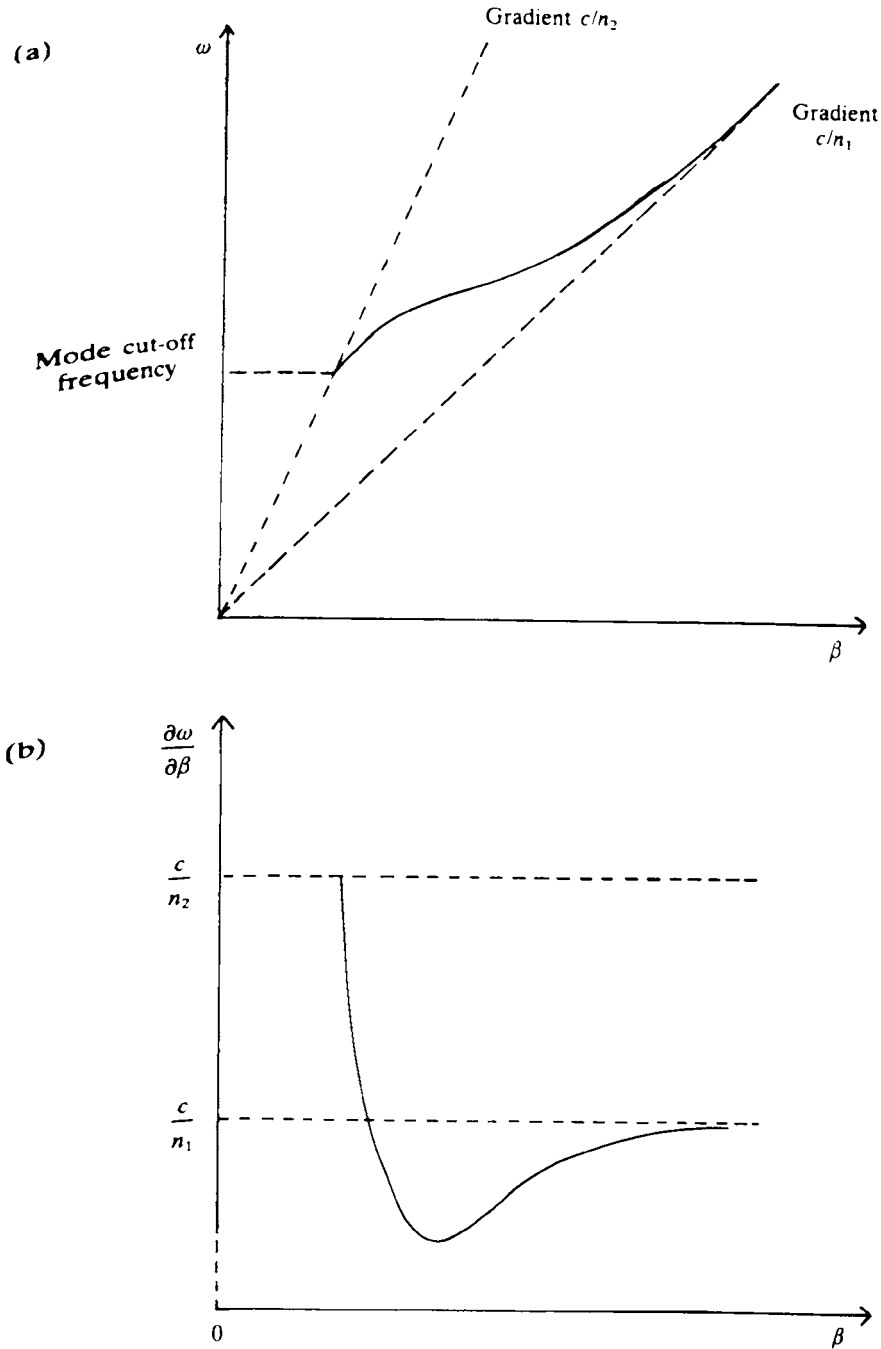


FIG. 8.17 Schematic relationship between (a) ω and β , and (b) between $\partial\omega/\partial\beta$ and β for a particular mode in a waveguide.

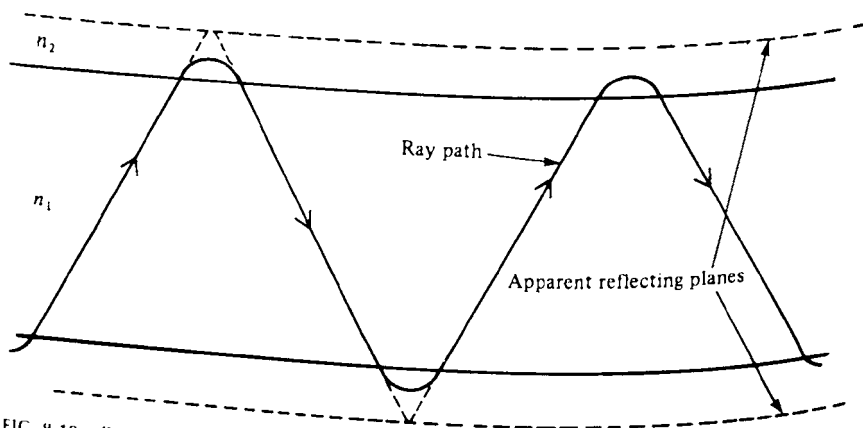


FIG. 8.18 Ray path in a waveguide illustrating the origin of the phase change on reflection in terms of penetration into the cladding. From within the guide, the ray behaves as if it were reflected from a plane a short distance inside the cladding.

HE modes, particularly those corresponding to highly 'skewed' rays, reach cut-off before an appreciable fraction of the modal energy extends into the cladding.

The values of the mode velocities at and far from cut-off also agree with what might be expected in view of the behaviour of the mode field distributions as shown in Fig. 8.12 for the planar waveguide. Far from cut-off, modes have most of their energy contained within the core and hence it seems reasonable that they should have a velocity corresponding to that of a plane wave travelling in a medium of refractive index n_1 . Modes close to cut-off, on the other hand, have an appreciable fraction of their mode energy in the cladding and hence might be expected to have velocities approaching c/n_2 .

When many modes are present, covering the whole range from close to cut-off to far from cut-off, then we expect that the mode velocities will have a spread of roughly $c/n_2 - c/n_1$, which is in fairly close agreement with the result of eq. (8.24) above.

EXAMPLE 8.6 Intermodal dispersion in step index fibers

Taking a step index fiber of length 1 km with $n_1 = 1.48$ and $n_2 = 1.46$ then we would expect a maximum pulse broadening due to intermodal dispersion as given by eq. (8.24) of

$$\Delta\tau^{\text{SI}} = \frac{1 \times 10^3}{3 \times 10^8} \frac{1.48}{1.46} (1.48 - 1.46) \text{ s}$$

that is,

$$\Delta\tau^{\text{SI}} = 6.76 \times 10^{-8} \text{ or } 68 \text{ ns}$$

Over long lengths of optical fiber (i.e. above 1 km or so), intermodal dispersion is found to be somewhat smaller than the value given by eq. (8.24). The reason for this is that a real

fiber is not perfectly uniform and also has small kinks or microbends which cause a phenomenon called *mode coupling*: that is, the energy in a particular mode may become transferred to another mode as a result of the radiation encountering one of these fiber deformations. This is illustrated in Fig. 8.19, where the possible effects of a sudden kink in a fiber are shown. Thus, although at a particular time a certain portion of the energy may be in a 'fast' mode and so be outdistancing the other modes, later on it may be in a slower mode thus enabling the rest to catch up. The outcome is that after an initial distance, usually about 1 km, over which dispersion is proportional to the length of fiber (as in eq. 8.24), the modes attain an 'equilibrium' relative energy content and the dispersion then becomes proportional to L^q where $q \approx 0.5$. A typical result is shown in Fig. 8.20.

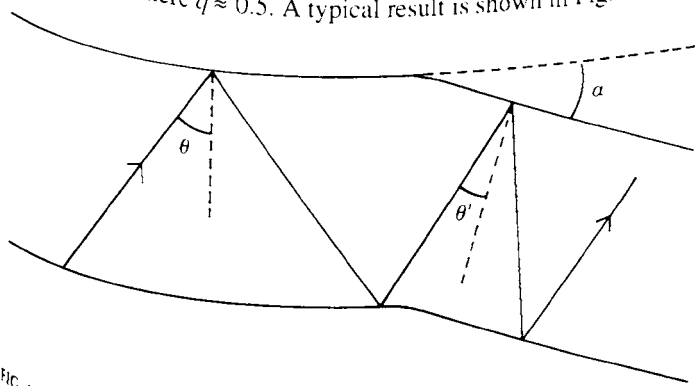


FIG. 8.19 In the presence of a small 'kink' of angle α in a fiber, the internal angles of the rays within the fiber will be altered. For the particular ray shown, θ changes to θ' where $\theta' = \theta - \alpha$.

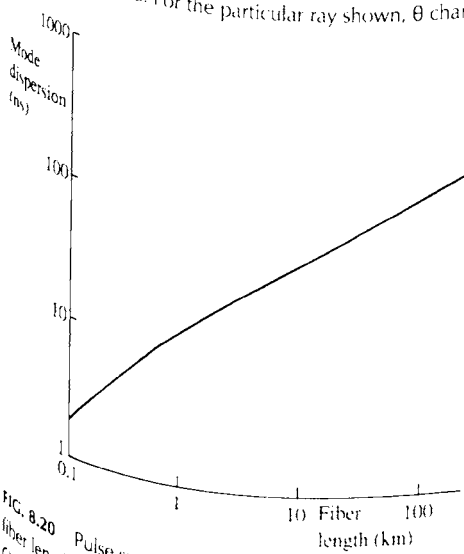


FIG. 8.20 Pulse spreading due to differences in mode velocity (intermodal dispersion) as a function of fiber length. For small lengths (up to about 1 km) the dispersion is proportional to the length traversed. Over larger distances the dispersion becomes approximately proportional to the square root of the length.

Despite the presence of mode coupling, the effects of intermodal dispersion still impose rather drastic limitations on the rate that information can be transmitted down the fiber. This will be discussed in more detail in Chapter 9; however, it is easy to appreciate at this stage that any narrow peaks in the optical irradiance will tend to broaden out in time as a function of distance along the fiber and hence reduce the information content (especially at high frequencies) of an amplitude-modulated signal. These unwanted consequences of intermodal dispersion may be alleviated either by designing a fiber where the mode velocities are more nearly equal than in step index fiber, or by using a fiber where only one mode is allowed to propagate (i.e. single mode fiber). As we shall see, single mode fibers have rather small diameter cores and it was not evident at first whether they could be manufactured reliably or indeed whether such problems as launching light into such fibers and joining them together could be readily solved. Thus it was that the first low dispersion fibers were of the so-called *graded index* type.

8.3.3 Graded index fibers

In a fiber where the refractive index in the core region is not constant but rather decreases smoothly from the centre to the cladding, the modes tend to travel with mode velocities which are closer together than in the case of step index fiber. The variation in refractive index with radial distance, r , in such fibers is often expressed in the form

$$\begin{aligned} n(r) &= n_1[1 - 2\Delta(r/a)^\alpha]^{1/2} & r < a \\ n(r) &= n_1(1 - 2\Delta)^{1/2} = n_2 & r \geq a \end{aligned} \tag{8.25}$$

where a is the core radius and Δ is defined in eq. (8.22). The parameter α determines the core refractive index variation (care must be taken not to confuse it with the absorption coefficient). Figure 8.21 shows the refractive index profile of a fiber where $\alpha = 2$.

The guided modes in such fibers are generally similar to those in step index fibers, and may also be referred to as LP_{lm} modes. In fact the step index profile can be seen to be a particular case of the above profile with $\alpha = \infty$. The number of guided modes, N , is given

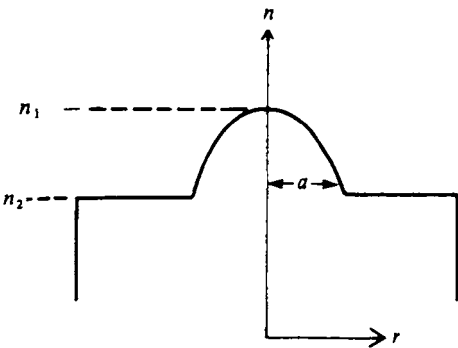


FIG. 8.21 Refractive index profile of a graded index fiber.

by (ref. 8.7)

$$N = \left(\frac{\alpha}{\alpha + 2} \right) a^2 \left(\frac{2\pi}{\lambda_0} \right)^2 n_1^2 \Delta$$

As expected this reduces to the step index result (eq. 8.20) when $\alpha = \infty$.

It may be shown (ref. 8.8) that intermodal dispersion is at a minimum when α is close to the value 2 and that with a monochromatic source the resulting minimum pulse broadening, $\Delta\tau_{\min}^{\text{GI}}$, over a distance L is given, approximately, by

$$\Delta\tau_{\min}^{\text{GI}} \approx \frac{Ln_1}{8c} \Delta^2 \quad (8.26)$$

Since we usually have $n_1 \approx n_2$, the result for the step index fiber dispersion (eq. 8.24) can be modified to

$$\Delta\tau^{\text{SI}} = \frac{L}{c} \frac{n_1}{n_2} (n_1 - n_2) \approx \frac{Ln_1}{c} \Delta$$

and so

$$\frac{\Delta\tau_{\min}^{\text{GI}}}{\Delta\tau^{\text{SI}}} \approx \frac{\Delta}{8} \ll 1$$

Thus the dispersion in optimally graded index fibers should be much smaller than in step index fibers (see Example 8.7). It is interesting to note that when $\alpha = 2$, the number of guided modes that the guide can support is half that for step index fiber of the same core radius. This implies that the amount of energy that can be coupled into graded index fiber is only about half of that which can be coupled into a comparable step index fiber.

EXAMPLE 8.7 Intermodal dispersion in graded index fibers

Taking an optimally graded fiber of length 1 km with the usual core and cladding refractive indices (i.e. $n_1 = 1.48$, $n_2 = 1.46$), we have that

$$\Delta = \frac{(1.48^2 - 1.46^2)}{2 \times 1.48^2} = 0.0134$$

From eq. (8.26) we have

$$\Delta\tau^{\text{GI}} \approx \frac{1 \times 10^3 \times 1.48}{8 \times 3 \times 10^8} (0.0134)^2 = 1.11 \times 10^{-10} \text{ s}$$

As expected this is a much smaller value for the dispersion than is the case in step index fibers (Example 8.6), where $\Delta\tau^{\text{SI}} = 6.76 \times 10^{-8} \text{ s}$.

Although in practice graded index fibers do indeed achieve much reduced values for the intermodal dispersion it is very difficult to obtain values that approach the predictions of

eq. (8.26). The reason for this lies in the extreme sensitivity of mode dispersion to variations in α , as shown in Fig. 8.22. Any slight variations in α that are likely to occur in fiber manufacture will easily increase the mode dispersion from its theoretical minimum value. The situation is further complicated by the presence of other forms of dispersion that become evident when intermodal dispersion is eliminated and which will be dealt with in section 8.3.6.

Before turning to single mode fibers it is interesting to analyze why, on a simple ray picture, graded index fibers have reduced intermodal dispersion. We consider a meridional ray which is travelling with a particular angle, θ , through the fiber centre. Instead of the fiber being continuously graded from core to cladding we suppose that the changes occur in a series of small steps. The ray will then undergo a series of refractions and, since the progression is from larger to smaller refractive indices, the value of θ will increase. Eventually θ will be large enough so that the ray, instead of being refracted, undergoes total internal reflection and is then directed back downwards towards the fiber centre as shown in Fig. 8.23(a). If we imagine that the refractive index steps increase in number whilst the change in refractive index at each step becomes smaller, then the ray path will tend towards the smooth sinusoid-like trajectory shown in Fig. 8.23(b). A more detailed analysis confirms this (ref. 8.9). At the optimum grading profile (i.e. where $\alpha \approx 2$) different modes have very nearly the same

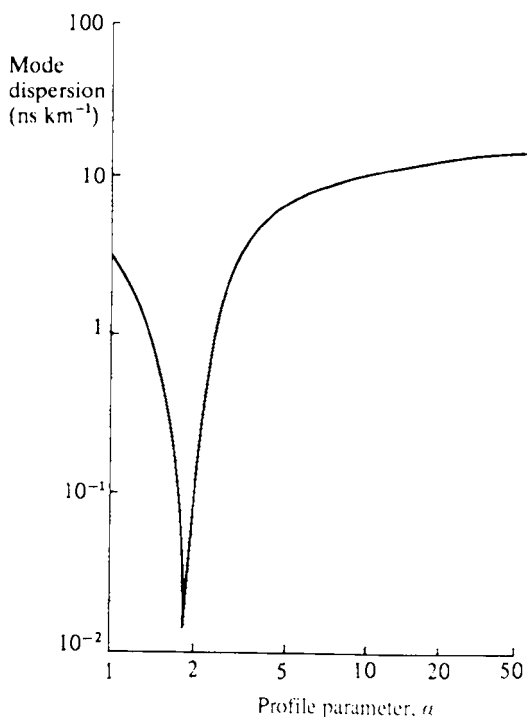


FIG. 8.22 Intermodal dispersion in a graded index fiber as a function of the profile parameter α . The curve has a very sharp minimum at a value of α just less than two.

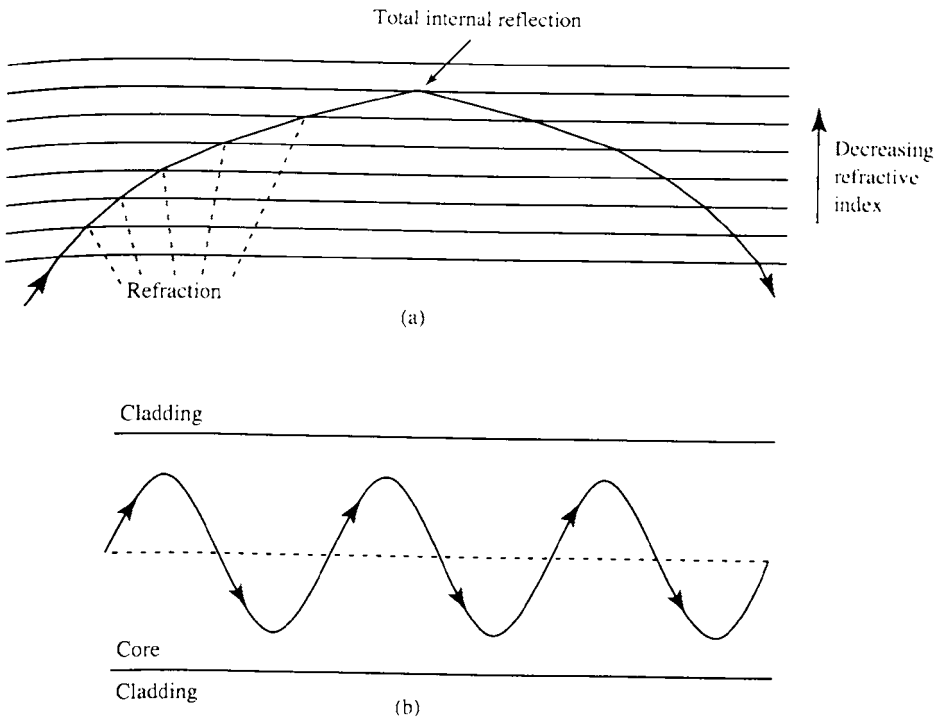


FIG. 8.23 (a) The path of a ray through a series of thin slabs of decreasing refractive index. (b) If the thickness of the slabs reduces to infinitesimally small values we obtain a sinusoid-like ray path.

‘wavelength’ along the guide, but different excursion ‘amplitudes’ away from the fiber axis (Fig. 8.24a).

We can now appreciate why mode dispersion is reduced in graded index fibers. Although the higher order modes (i.e. those with larger ‘amplitudes’) have longer path lengths than do the lower order modes, when they are further away from the centre of the fiber they are travelling in regions of lower refractive index, and hence higher ray velocity. Thus the higher order modes can compensate to some extent for their longer paths by having a higher average velocity.

Only meridional-type rays have been discussed so far, yet it should be mentioned that the equivalents of skew rays in graded index fibers also exist and, perhaps not very surprisingly, turn out to be ‘helical’ rays which avoid the centre (Fig. 8.24b).

8.3.4 Single mode fibers

We have already mentioned that when the fiber V parameter is less than 2.405 then only one mode (the HE_{11} mode, or, in the linearly polarized approximation, the LP_{01} mode) can propagate. Actually, strictly speaking *two* HE_{11} modes can be present with orthogonal

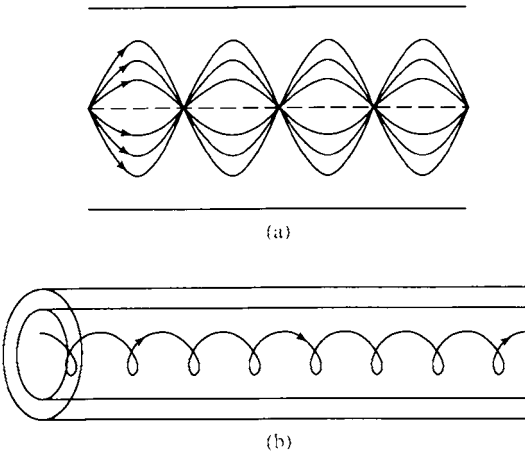


FIG. 8.24 Ray paths in a graded index fiber for (a) meridional rays and (b) helical rays which avoid the centre.

polarizations, but for simplicity we will assume that we are dealing with only one of these. In theory, the HE_{11} mode will propagate no matter how small the value of V . As V decreases, however, the mode field will extend increasingly into the cladding (see the discussion following eq. 8.18); if the field then becomes at all significant at the edge of the cladding, appreciable amounts of energy may be lost from the fiber, leading to the mode being highly attenuated.

In terms of the fiber core radius, a , the condition (eq. 8.19) for a single mode to propagate is

$$a < \frac{2.405\lambda_0}{2\pi(n_1^2 - n_2^2)^{1/2}}$$

or

$$a < \frac{2.405\lambda_0}{2\pi(\text{NA})} \tag{8.27}$$

This relationship implies that single mode fibers will have cores that are only of the order of λ_0 (i.e. micrometres) in radius. It is advantageous from a number of points of view, however, that they have as large a diameter as possible. From eq. (8.27) we can see that this may be done by reducing the NA value, that is by making the core and cladding refractive indices very close together. In practical terms single mode fibers are made with NA values of the order of 0.1, with a typical design criterion for a single mode fiber being $2 \leq V \leq 2.2$. When used with radiation in the wavelength region $1.3 \mu\text{m}$ to $1.6 \mu\text{m}$, single mode fibers have core diameters that are typically between $5 \mu\text{m}$ and $10 \mu\text{m}$ as illustrated in Example 8.8.

EXAMPLE 8.8 Single mode fiber radius

If we take our usual fiber core and cladding refractive indices (i.e. $n_1 = 1.48$ and $n_2 = 1.46$), this results in a numerical aperture of 0.242. If radiation of $1.5 \mu\text{m}$ is being used then from eq. (8.27) the maximum core radius for single mode operation is given by

$$a < \frac{2.405 \times 1.5 \times 10^{-6}}{2\pi \times 0.242}$$

that is,

$$a < 2.37 \mu\text{m}$$

With an NA value of 0.1, the maximum value of a is increased to $5.74 \mu\text{m}$. With a design criterion of $V = 2$ the core diameter would then be $9.5 \mu\text{m}$.

It should be noted that for a given core diameter a particular fiber will only be single mode when the wavelength of radiation being used is greater than a critical value, λ_c , which is called the *cut-off wavelength* (since it represents the wavelength at which the mode above the lowest order mode cuts off). From eq. (8.27) we have that

$$\lambda_c = 2\pi a(\text{NA})/2.405 \quad (8.28)$$

When a fiber is to be used as a single mode fiber, care must be taken to ensure that the wavelengths used never exceed the cut-off wavelength.

Although the mode field distribution in a single mode fiber is theoretically described by Bessel functions, it is convenient to represent the field irradiance distribution (with little loss in accuracy) by the much simpler Gaussian function, that is

$$I(r) = I_0 \exp(-2r^2/\omega_0^2) \quad (8.29)$$

where $2\omega_0$ is known as the *mode field diameter*. ω_0 thus represents the radial distance at which the mode irradiance has fallen to $\exp(-2)$ (i.e. 13.5%) of its peak value. As the V parameter of a fiber gets smaller we would expect that ω_0 will increase (i.e. the field will extend further into the cladding). A useful empirical relationship between ω_0 and V which is accurate to better than 1% if $1.2 < V < 3$ is given by (ref. 8.10)

$$\omega_0/a = 0.65 + 1.619V^{-3/2} + 2.879V^{-6} \quad (8.29a)$$

EXAMPLE 8.9 Mode field diameters in single mode fiber

A typical design criterion for a single mode fiber is $V = 2$. Under these circumstances the mode field irradiance diameter is given by eq. (8.29a) as

$$\omega_0/a = 0.65 + 1.619 \times (2)^{-3/2} + 2.879 \times (2)^{-6} = 1.049$$

Thus in this particular case the mode field radius is slightly greater than the core radius. The resulting mode irradiance distribution is shown in Fig. 8.25.

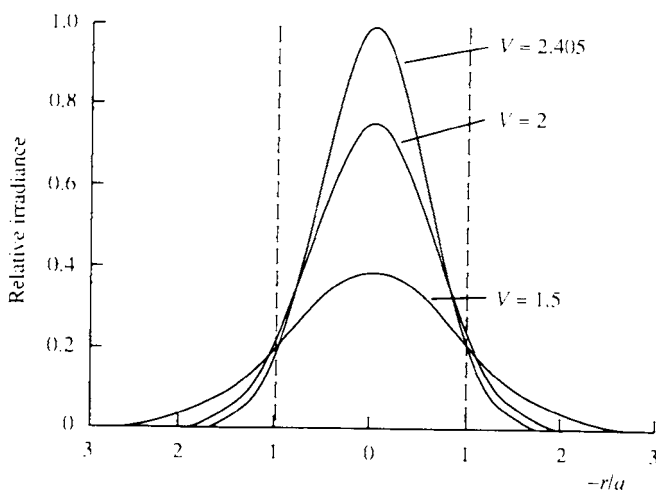


FIG. 8.25 The radial variation of the mode field in a single mode fiber. As the V parameter decreases the mode field extends further and further into the cladding.

8.3.5 Fiber materials and types

At this stage it is useful to mention briefly the types of fiber that are available, the materials from which they are made and typical dimensions. For a fiber material to be of any practical use it must obviously be highly transparent to the radiation being used. It must also be possible to fabricate the basic core and cladding structures as discussed above and finally it must be reasonably flexible so that it can be taken round bends etc. There are in fact only two materials that satisfy these criteria and which currently are widely used in fiber manufacture, these being silica (SiO_2) and various types of plastic.

Typical plastic fibers have core and cladding made from different types of plastic. They have core diameters of about $980\text{ }\mu\text{m}$ and core + cladding diameters of about $1000\text{ }\mu\text{m}$. Maximum diameters are limited to a few millimetres, beyond which the fibers start to become rather inflexible. The two different types of plastic used for the core and cladding can have significantly different refractive indices, resulting in relatively large numerical apertures (i.e. 0.4 or greater). The main problem with plastic fibers is that they exhibit much higher attenuations than silica fibers. A more detailed discussion of plastic fibers is contained in section 8.7.3.

Silica fibers by contrast can be made with attenuations as low as 0.2 dB km^{-1} (see section 8.4.2). However, the fibers become increasingly inflexible when their diameters exceed $600\text{ }\mu\text{m}$; typical multimode step index fibers have core diameters of $200\text{ }\mu\text{m}$, with core + cladding diameters of $250\text{ }\mu\text{m}$. The necessary differences between the core and cladding refractive indices are obtained by doping the silica. For example, doping silica with germania (GeO_2) raises the refractive index, so that a silica-based fiber could be made from a germania core with a pure silica cladding. Since it is difficult to change the refractive index of silica by more than a few per cent by doping, the fibers have rather smaller

numerical apertures (about 0.2 or less) than do plastic fibers. Both graded index and single mode silica fibers are readily available. The former have rather smaller core diameters than step index fibers (usually 50 μm or 62.5 μm). As has been mentioned previously, almost all silica fibers have a coating of plastic material round the cladding. This serves no optical purpose but is vital in maintaining the strength of the fiber and preventing physical and chemical damage. The overall diameter of the fiber may then increase to between 400 μm and 500 μm .

8.3.6 Dispersion in single mode fibers

By using single mode fiber the problems associated with intermodal dispersion are completely avoided. However, there still remain three other sources of dispersion, namely *material* (or *chromatic*) *dispersion*, *profile dispersion* and *waveguide dispersion*.

Material dispersion arises because of the wavelength dependence of the refractive index of the fiber material and applies equally to a plane wave travelling in a medium of infinite extent as well as in waveguides. We saw in section 1.2 (eq. 1.8) that a pulse of radiation consisting of a finite spread of wavelengths travels with the *group velocity* v_g given by

$$v_g = \frac{d\omega}{dk}$$

Here $\omega = 2\pi\nu$ and $k = 2\pi/\lambda$ (λ being the wavelength in the material); thus

$$v_g = \frac{dv}{d(1/\lambda)} = -\lambda^2 \frac{dv}{d\lambda} \quad (8.30)$$

It is convenient to work with the vacuum wavelength λ_0 , where $n\lambda = \lambda_0$, n being the refractive index of the medium.

Now

$$\frac{dv}{d\lambda} = \frac{dv}{d\lambda_0} \frac{d\lambda_0}{d\lambda} = \frac{dv}{d\lambda_0} \left(\frac{1}{n} - \frac{\lambda_0}{n^2} \frac{dn}{d\lambda_0} \right)^{-1}$$

so that

$$v_g = \frac{c}{(n - \lambda_0 \frac{dn}{d\lambda_0})} \quad (8.31)$$

A parameter that may be usefully introduced at this stage for later use is the material *group index* N_g , where

$$N_g = c/v_g$$

Thus

$$N_g = n - \lambda_0 \frac{dn}{d\lambda_0} \quad (8.32)$$

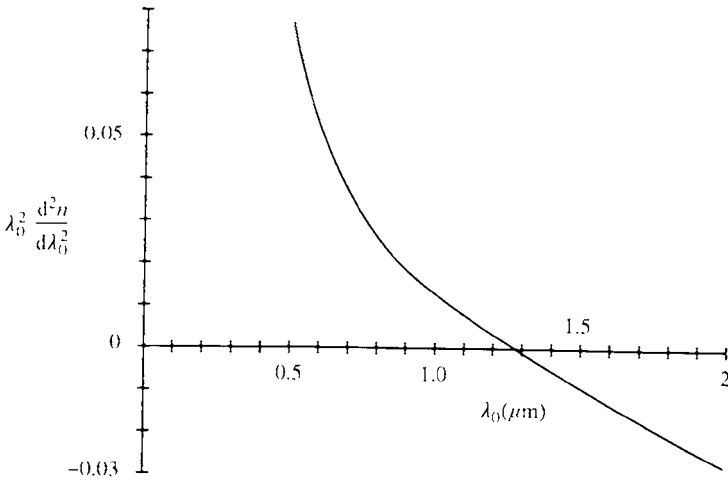


FIG. 8.26 Variation of the quantity $\lambda_0^2 d^2n/d\lambda_0^2$ with wavelength for pure silica (SiO_2). The function becomes zero at about $1.3 \mu\text{m}$, which results in very small values for the material dispersion near this wavelength. The addition of dopants displaces the curve slightly.

Now the group transit time, τ_g , for radiation of wavelength λ_0 to travel a distance L in the medium is given by

$$\begin{aligned} \tau_g &= L/v_g \\ \tau_g &= \frac{L}{c} \left(n - \lambda_0 \frac{dn}{d\lambda_0} \right) \end{aligned} \tag{8.33}$$

If the radiation in fact has a finite wavelength spread $\Delta\lambda_0$, then the resulting spread in transit times over the length L will be given by $\Delta\tau^{\text{mat}}$ where

$$\begin{aligned} \Delta\tau^{\text{mat}} &\approx \frac{d\tau_g}{d\lambda_0} \Delta\lambda_0 \\ \Delta\tau^{\text{mat}} &\approx \frac{L\lambda_0^2}{c} \frac{d^2n}{d\lambda_0^2} \frac{\Delta\lambda_0}{\lambda_0} \end{aligned} \tag{8.34}$$

As far as pulse spreading is concerned we are usually only interested in the magnitude of $\Delta\tau^{\text{mat}}$, although the sign of $\Delta\tau^{\text{mat}}$ in fact determines whether the long or short wavelengths arrive first (this may sometimes be important, see e.g. section 9.3.10 on solitons).

A graph of the dimensionless quantity $\lambda_0^2 d^2n/d\lambda_0^2$ as a function of wavelength for silica is shown in Fig. 8.26 and calculations of material dispersion based on this are given in Example 8.10.

EXAMPLE 8.10 Material dispersion for a laser and LED source _____

We use eq. (8.34) to estimate material dispersion effects over 1 km of silica-based fiber

when the following two sources are used:

- (a) a GaAs LED operating at 850 nm and with a 50 nm linewidth; and
- (b) a semiconductor laser operating at 1550 nm and with a linewidth of 3 nm.

From Fig. 8.26 we have that at 850 nm, $\lambda_0^2 d^2n/d\lambda_0^2 = 2.14 \times 10^{-2}$, whilst at 1550 nm

$$\lambda_0^2 d^2n/d\lambda_0^2 = -1.02 \times 10^{-2}$$

Using eq. (8.34) we then have

$$\begin{aligned} \text{(a) for the LED} \quad \Delta\tau^{\text{mat}} &= \frac{10^3}{3 \times 10^8} 2.14 \times 10^{-2} \frac{50 \times 10^{-9}}{850 \times 10^{-9}} \\ &= 4.2 \times 10^{-9} \text{ s} \end{aligned}$$

$$\begin{aligned} \text{(b) for the laser} \quad \Delta\tau^{\text{mat}} &= \frac{10^3}{3 \times 10^8} 1.02 \times 10^{-2} \frac{3 \times 10^{-9}}{1550 \times 10^{-9}} \\ &= 6.6 \times 10^{-11} \text{ s} \end{aligned}$$

The superiority of the laser source operating at 1550 nm is at once apparent.

Inspection of Fig. 8.26 shows that just below $1.3 \mu\text{m}$ the value of $\lambda_0^2 d^2n/d\lambda_0^2$ becomes zero. We denote the wavelength at which this occurs by λ_{min} . Unfortunately it is not the case that material dispersion vanishes altogether when the central wavelength of the source is equal to λ_{min} . Although the longest and shortest source wavelengths will be travelling with the same group velocity, the central wavelength in the group will have a slightly higher velocity. The situation is dealt with in Problem 8.12. It is true, however, that material dispersion will become extremely small at, or close to, λ_{min} .

We must now deal with the two remaining sources of dispersion in single mode optical fibers. Profile dispersion arises because of the variation of the quantity Δ with wavelength. In other words, it depends on the fact that the changes in refractive indices for core and cladding with wavelength may not be the same. In practice, since in many fibers both core and cladding are derived from the same material (i.e. silica) the dependence of Δ (as defined by eq. 8.22) on wavelength is very small and for all practical purposes we may neglect profile dispersion. Waveguide dispersion, however, cannot be neglected. It arises because the mode propagation velocity itself depends on wavelength regardless of any refractive index variations of the medium. To appreciate how this comes about we may refer back to the section on propagation in planar dielectric waveguides (section 8.2). From Fig. 8.8 we can see that if the wavelength propagating in the guide is increased slightly then the value of the expression $(2\pi dn_1 \cos \theta)/\lambda_0$ will decrease and hence the points of intersection with the curves of $y = m\pi + \phi(\theta)$ and the curve $y = (2\pi dn_1 \cos \theta)/\lambda_0$ will occur at smaller values of θ . This in turn will change the effective velocity of the ray down the waveguide (in fact, on a simple ray model the effective velocity will decrease as λ_0 increases).

For a step index single mode fiber waveguide dispersion is negative and becomes increasingly negative with increasing wavelength. A diagram illustrating both the material and waveguide dispersion together with the resulting total dispersion as a function of wavelength is

given in Fig. 8.27. In general the minimum dispersion wavelength for the total dispersion occurs at a somewhat higher wavelength than that for which material dispersion is a minimum. Since, in fact, the waveguide dispersion increases with decreasing values of the parameter V , reducing the core diameter will give an increased waveguide dispersion, and will thus shift λ_{\min} to higher wavelengths. The actual value of λ_{\min} depends on a , Δ and also the details of the refractive index profile. By a suitable choice of these parameters it is possible to make λ_{\min} lie anywhere between $1.3\ \mu\text{m}$ and $1.6\ \mu\text{m}$. Such fibers are referred to as *dispersion shifted* fibers. By using more complicated fiber structures such as the so-called *depressed cladding* fiber (Fig. 8.28), it is possible also to obtain *dispersion flattened* fibers where the dispersion curve crosses the zero axis at *two* wavelengths (Fig. 8.29). This gives

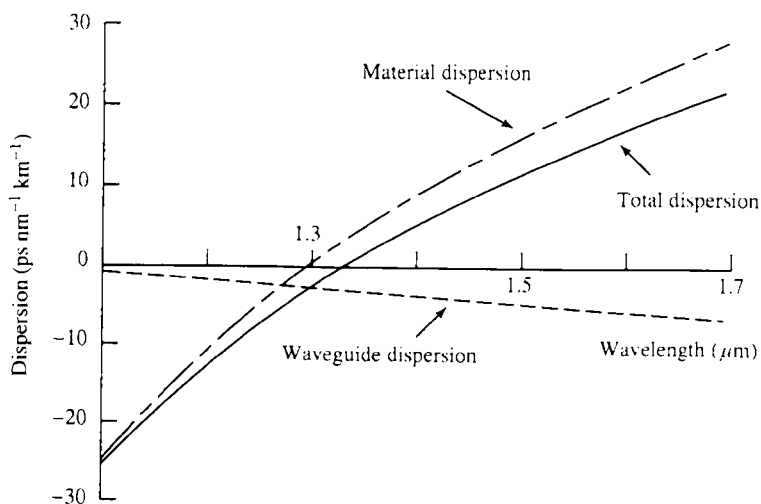


FIG. 8.27 Illustration of the magnitudes of material dispersion together with waveguide dispersion as a function of wavelength; the resultant dispersion is also shown.

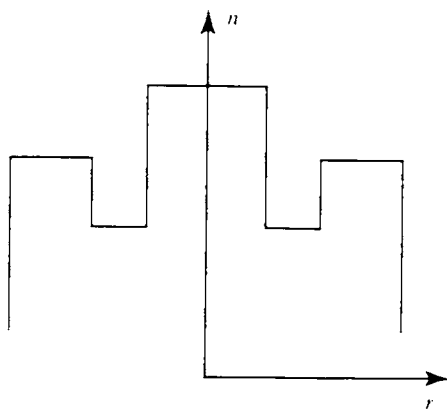


FIG. 8.28 Refractive index profile of the 'W' (or depressed cladding) fiber.

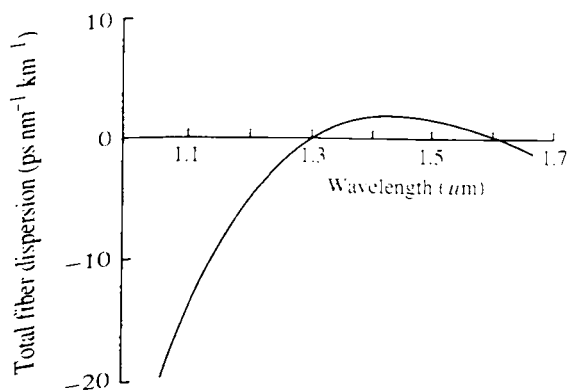


FIG. 8.29 The variation of total dispersion with wavelength in a dispersion-flattened fiber.

rise to a fiber with relatively low dispersion over an appreciable wavelength range, which is useful since, as we shall see in section 8.4, silica-based fibers exhibit minimum losses at a wavelength of $1.55 \mu\text{m}$. Thus it is possible to manufacture fibers which exhibit a minimum dispersion and loss at the same wavelength.

8.4

Losses in fibers

We may divide fiber losses into two categories: (a) those which result from the distortion of the fiber from the ideal straight line configuration (*extrinsic* losses) and (b) those that are inherent in the fiber itself (*intrinsic* losses).

8.4.1 Bending losses

In multimode fibers the principal bending loss mechanism arises from mode coupling as discussed in section 8.3.2. From there we remember that bends in a multimode fiber can lead to the possibility that energy in a particular mode is coupled into a different mode. If the initial mode is one that is close to cut-off, then, after the coupling process, the energy might find itself in a non-guided mode (i.e. with $\theta < \theta_c$) which is refracted out of the core and into the cladding. It is evident that the energy which is in higher order modes is more likely to be lost in this way than that in the lower order modes.

When a fiber is bent into an arc of a circle, the loss is found to be strongly dependent on the radius of curvature, and a semiempirical expression that is sometimes used to describe the loss is

$$\alpha_B = C \exp(-R/R_c) \quad (8.35)$$

where α_B is the absorption coefficient due to the bend, R the bend radius and R_c a constant which depends on the fiber parameters. Quite considerable losses can be observed when bend

radii of the order of millimetres are involved, although such excessive deformations are quite easily avoided in practice and in any case may very well cause fracture of the fiber.

Another cause of loss is a continuous succession of small deformations or *microbends* which can occur when the fiber is pressed against a surface which is not perfectly smooth (Fig. 8.30). A detailed analysis of the situation when a fiber is deformed into a small amplitude sinusoid of wavelength Λ (ref. 8.11) shows that modes with propagation constants β_1 and β_2 will be strongly coupled together provided that

$$\Lambda^{-1} = \frac{|\beta_1 - \beta_2|}{2\pi} \quad (8.36)$$

This expression shows that higher order modes are coupled together with relatively small values of Λ , whilst adjacent lower order modes are coupled with larger Λ values. Here 'small' means typically the order of 1 mm, whilst 'large' means about 20 mm or more. Since it is the higher order modes that are most susceptible to mode coupling losses, periodic deformations with a periodicity of the order of a few millimetres are best avoided. Such microbends can be caused by bare fiber being pressed against a protective jacket or wound tightly on a drum. Unless some care is taken, microbending loss can add appreciably to the intrinsic loss in a fiber (ref. 8.12).

In single mode fibers bending loss is also present but a different mechanism is responsible. Figure 8.31 shows the field within a single mode fiber as it propagates round a bend. It can be seen that the part of the mode which is on the outside of the bend will need to travel faster than the part on the inside to maintain a wavefront that is perpendicular to the propagation direction. Now each mode extends, in theory, an infinite distance into the cladding despite the exponential decline of the electric field within it (see eq. 8.8). Consequently some part of the mode in the cladding will find itself attempting to travel at greater than the speed of light. Since, according to Einstein's theory of relativity, this is not possible, the energy associated with this particular part of the mode must be radiated away. It is reasonable to deduce that the loss will be greater (a) for bends with smaller radii of curvature and (b) when the mode extends most into the cladding. In fact the loss can be represented by a function similar to that of eq. (8.35) with R_c being equal to $a/(\text{NA})^2$ (ref. 8.13) where a is the fiber radius. Bend radii of the order of R_c are likely to result in considerable losses, although in fact the mechanical properties of the fiber usually cause the fiber to break long before such small bend curvatures can be achieved.

8.4.2 Intrinsic fiber losses

In this section we examine the losses in silica-based fiber that depend on the fiber material



FIG. 8.30 Small kinks or 'microbends' can be formed when a fiber is pressed against a surface that is not perfectly smooth.

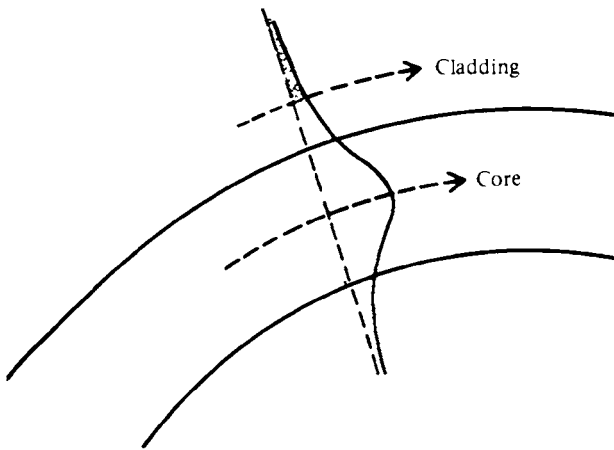


FIG. 8.31 Illustration of the mechanism of radiation loss in fibers at bends. To maintain a plane wavefront, a part of the mode (shown shaded on the diagram) may have to travel at velocities greater than that of light in the cladding. Since this is not possible, this portion of the mode must be radiated away.

itself (plastic fibers are discussed in section 8.7.3). Before discussing details of the loss mechanisms it is useful to consider the unit commonly used to characterize loss in fibers. This is dB km^{-1} , and is defined as follows: if a beam of optical power P_i is launched into one end of a fiber and if, after a length L of fiber, the power remaining in the fiber is P_r , then the loss (or attenuation) is given by

$$\text{attenuation} = \frac{10 \log_{10}(P_i/P_r)}{L} \text{ dB km}^{-1} \quad (8.37)$$

Losses intrinsic to silica fibers have two main sources: (a) scattering losses and (b) absorption losses.

SCATTERING LOSSES

We have assumed in our discussion of light propagation in fibers that the material is homogeneous. Silica is an amorphous material and thus suffers from structural disorder: that is, the same basic molecular units are present throughout the material but these are connected together in an essentially random way. This results in a fluctuation in the refractive index through the material with each irregularity acting as a point scattering centre. The scale of the fluctuations is of the order of $\lambda/10$ or less, and the scattering is known as *Rayleigh scattering* and characterized by an absorption coefficient that varies as λ^{-4} (ref. 8.14). In detail the absorption coefficient due to Rayleigh scattering (α_R) can be written

$$\alpha_R = \frac{8\pi^3}{3\lambda_0^4} n^8 p^2 \beta k T_F \quad (8.38)$$

where n is the refractive index, p the photoelastic coefficient and β the isothermal compressibility at the temperature T_F , which itself is the temperature at which the disorder becomes effectively 'frozen in' as the glass cools.

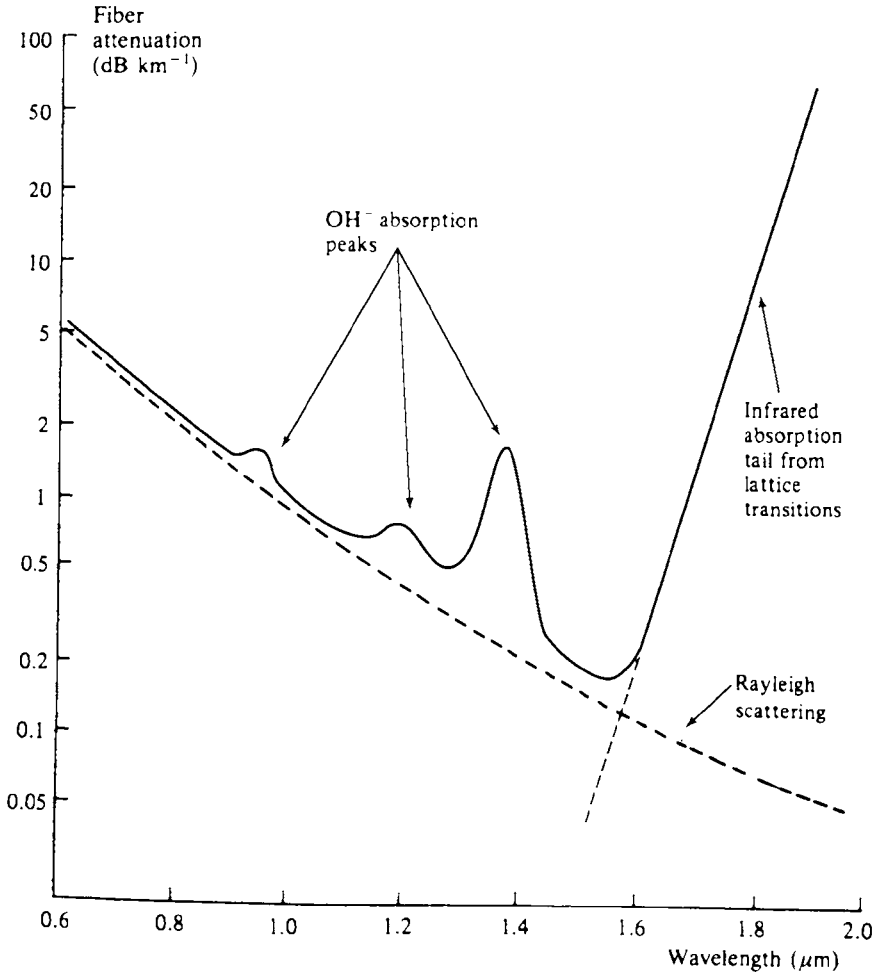


FIG. 8.32 Typical attenuation versus wavelength plot for a silica-based optical fiber. The contribution from Rayleigh scattering is shown, as are the other two main loss mechanisms, namely the infrared absorption tail and the hydroxyl (OH^-) absorption peaks.

EXAMPLE 8.11 Rayleigh scattering loss

We estimate the contribution that Rayleigh scattering makes to the attenuation in a silica fiber at a wavelength of $1\ \mu\text{m}$. Taking the values $n=1.45$, $p=0.286$, $\beta=7 \times 10^{-11}\ \text{m}^2\text{N}^{-1}$ and $T_F=1400\ \text{K}$, we obtain from eq. (8.38)

$$\begin{aligned}\alpha_R &= \frac{8\pi^3}{3(10^{-6})^4} \times (1.45)^8 \times 7 \times 10^{-11} \times 1.38 \times 10^{-23} \times 1400 \\ &= 1.79 \times 10^{-4}\ \text{m}^{-1}\end{aligned}$$

Over a length L of fiber, an absorption coefficient of α_R will give rise to a loss of

$-10 \log_{10}[\exp(-\alpha_R L)]$ dB. Over a kilometre of fiber, we should then expect a loss of $-10 \log_{10}[\exp(-1.79 \times 10^{-4} \times 10^3)]$ dB km⁻¹ or 0.78 dB km⁻¹.

Lower attenuations will be obtained at longer wavelengths. At 1.55 μm , for example (where fiber attenuation is at a minimum), we would expect the absorption to be a factor 1.55⁴ or 5.8 smaller, that is $\alpha_R = 3.1 \times 10^{-5} \text{ m}^{-1}$ which is equivalent to 0.13 dB km⁻¹.

ABSORPTION LOSSES

Absorption losses in the visible and near-infrared regions arise mainly from the presence of impurities, particularly traces of transition metal ions (e.g. Fe³⁺, Cu²⁺) or hydroxyl ions (OH⁻). The latter are responsible for absorption peaks at 0.95 μm , 1.24 μm and 1.39 μm . Most of the dramatic successes in reducing fiber losses came about because of better control of impurity concentrations.

At wavelengths greater than about 1.6 μm , the main losses are due to transitions between vibrational states of the lattice itself. Although the actual fundamental absorption peaks occur at wavelengths well into the infrared (in SiO₂, for example, the main peak is at 9 μm), the anharmonic nature of the interatomic forces gives rise to combination and overtone bands; that is, an incoming photon can simultaneously excite two or more fundamental lattice vibrations (or phonons). Thus, a number of strong absorption bands extend all the way down to about 3 μm with appreciable absorption still occurring below 2 μm . To date, some of the lowest attenuations have been obtained with GeO₂-doped SiO₂ fibers (ref. 8.15), which have shown minimum attenuations of about 0.2 dB km⁻¹ at 1.55 μm . This figure is quite close to the limit set by Rayleigh scattering (see Example 8.11). A typical attenuation curve is shown in Fig. 8.32.

8.5

Optical fiber connectors

8.5.1 Single fiber jointing

It is obviously impractical to assume that the correct lengths of fiber will always be available for every situation and hence we need to be able to join fibers together in a manner which gives rise to minimum light loss. As well as permanent connections there may also be occasions when demountable connections are required.

Perhaps the simplest technique for permanently joining fibers together is to use a *fusion splice*. Here (Fig. 8.33) the fiber ends are held in close proximity and heated up to the material softening point (about 2000°C for silica). The fibers are then pushed together and the two ends fuse to form a continuous length of fiber. Any slight initial misalignment of the fiber ends tends to be corrected as surface tension forces pull the fibers back into alignment. The finished join will be mechanically weak since the protective coating layer has to be removed before the process starts. To provide protection it is usual to slide a small metal tube over the join and seal it onto the fiber using epoxy resin. Both multimode and single mode fibers can be joined by this means and transmission losses can be less than 0.1 dB.

Many other types of jointing, both permanent and demountable, involve a *butt join* whereby the two fiber ends are held in close proximity. There are two types of loss associated with

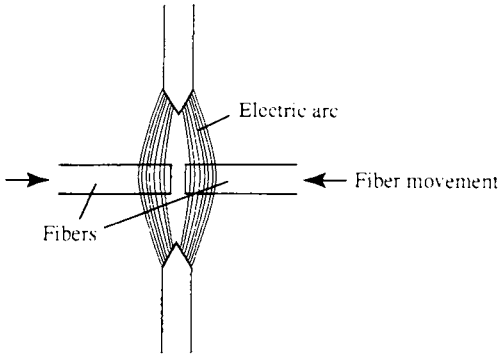


FIG. 8.33 In fusion splicing the fiber ends are pushed together whilst being heated within an electric arc.

this: Fresnel losses, caused by back reflection at the silica/air interfaces, and losses associated with any misalignment of the end faces. The Fresnel losses are readily estimated: if we assume light hits the fiber end at normal incidence then the fraction of light, R_F , reflected back at each fiber end is given by

$$R_F = \left(\frac{n_1 - n_0}{n_1 + n_0} \right)^2 \quad (8.39)$$

where n_0 is the refractive index of the medium between the fibers and n_1 that of the fiber core. The total transmission for each face due to Fresnel reflection is then given by

$$T_F = 1 - R_F \quad (8.39a)$$

Since we have two interfaces the total loss will then be $(R_F)^2$. Usually this amounts to a few tenths of a decibel (Example 8.12).

EXAMPLE 8.12 Fresnel losses at fiber joins

If we take a fiber core with refractive index 1.48 and assume that the medium between the fibers is air ($n_0 = 1$), then R_F is given by

$$R_F = \left(\frac{1.48 - 1}{1.48 + 1} \right)^2 = 0.0375$$

Thus $T_F = (1 - 0.0375)^2$ or 0.926. This corresponds to a loss of $-10 \log_{10}(0.926)$ or 0.33 dB.

The second type of loss associated with butt-type connections arises from misalignment between the fibers. There are three basic types of misalignment to be considered: (a) the distance between the fibers along the fiber axis (*longitudinal*); (b) the offset distance perpendicular to the fiber axis (*lateral*); and (c) the angle between the two axes of the fibers (*angular*). These are illustrated in Fig. 8.34. For both multimode and single mode fibers the

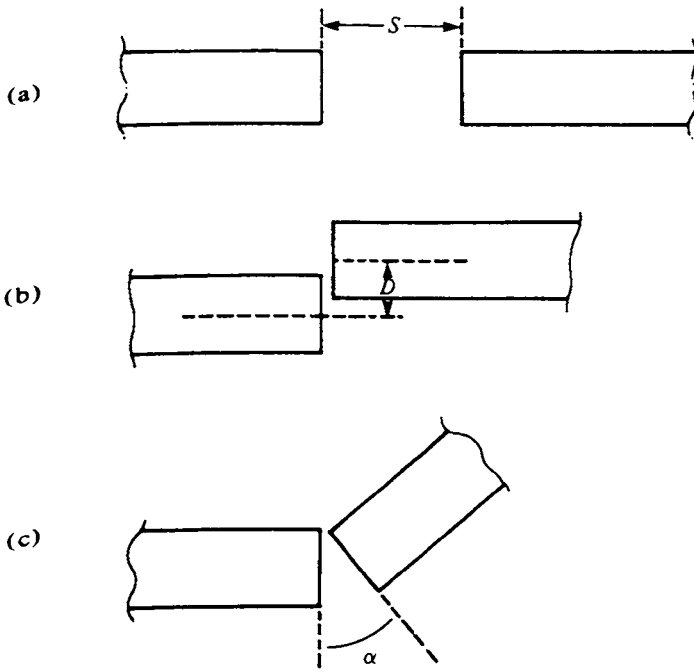


FIG. 8.34 Illustration of the three types of misalignment mentioned in the text: (a) longitudinal, (b) lateral and (c) angular misalignments. The parameters used to describe these (S , D and α) are also shown.

most important of these is usually lateral loss. If we assume that step index multimode fibers have a uniform irradiance across the fiber ends then it is fairly easy from a simple determination of the end face overlap area to show that the transmission losses resulting from lateral misalignment can be written (see Problem 8.17)

$$T_{\text{lat}} = \frac{2}{\pi} \left\{ \cos^{-1} \left(\frac{D}{2a} \right) - \frac{D}{2a} \left[1 - \left(\frac{D}{2a} \right)^2 \right]^{1/2} \right\} \quad (8.40)$$

where D is the lateral displacement and a the fiber core radius. Figure 8.35 shows typical results for lateral and longitudinal misalignments in step index multimode fiber.

The corresponding results for single mode fibers (and indeed for graded index fibers) tend to be rather more difficult to obtain; for a summary the interested reader is referred to ref. 8.16.

EXAMPLE 8.13 Transmission loss from lateral misalignment

We consider a multimode step index fiber where we have a lateral misalignment which amounts to some 10% of the fiber diameter. That is, $D/2a = 0.1$. From eq. (8.40) we have

$$T_{\text{lat}} = \frac{2}{\pi} \{ \cos^{-1}(0.1) - 0.1[1 - (0.1)^2]^{1/2} \} = 0.872$$

The transmission loss is therefore $-10 \log_{10}(0.872) = 0.59$ dB.

This calculation excludes any Fresnel losses (Example 8.12); if these are included the total loss becomes $0.59 + 0.33 = 0.92$ dB.

As far as making permanent joints, which rely on butting fibers together, is concerned there are three main steps. First the two end faces of the fibers are made as flat and as perpendicular to the fiber axis as possible. Secondly the fibers are held in some sort of jig which allows them to be closely aligned. An example of this is the V-groove arrangement of Fig. 8.36; another possibility is to put the fibers inside a closely fitting capillary tube. The final step is to fix the fibers permanently in place, usually by the use of an epoxy-resin-type glue.

When it comes to butt joining demountable connectors, there is a wide variety of different types available from various manufacturers. One example is the biconical type illustrated in Fig. 8.37; other types are described in ref. 8.17. With all types of butt joins it is possible to reduce the contribution from Fresnel loss by filling the gaps between the fiber ends with a fluid or grease which has a refractive index as close as possible to that of the fiber core. However, in practical jointing situations the problem of dirt being picked up in the fluid often outweighs the potential advantages.

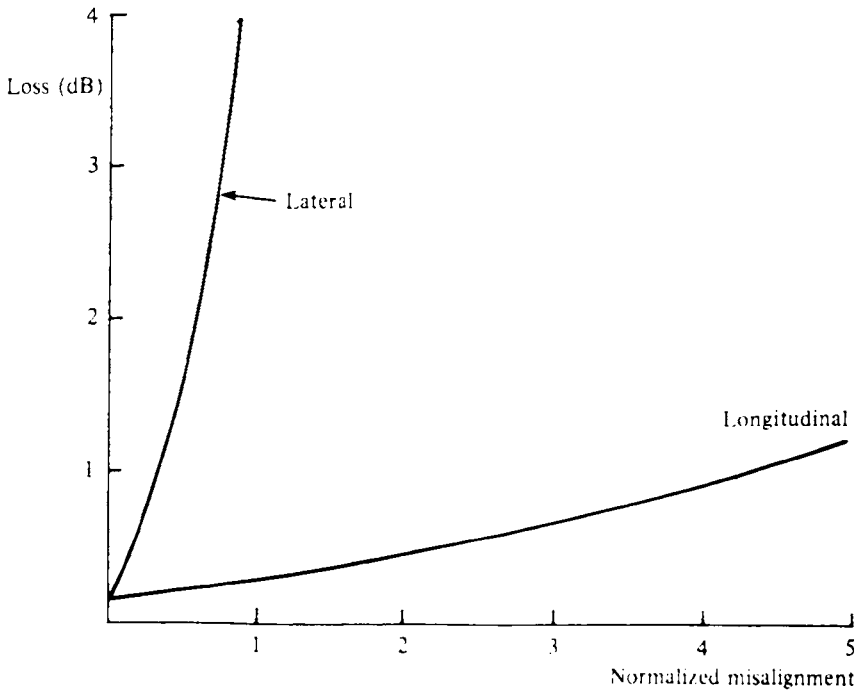


FIG. 8.35 Effects of separation and offset misalignments on fiber connection losses in multimode fibers. The misalignments are expressed in terms of the normalized parameters S/a (longitudinal) and D/a (lateral), where a is the fiber core radius. The residual loss when the misalignment is reduced to zero is due to Fresnel losses.

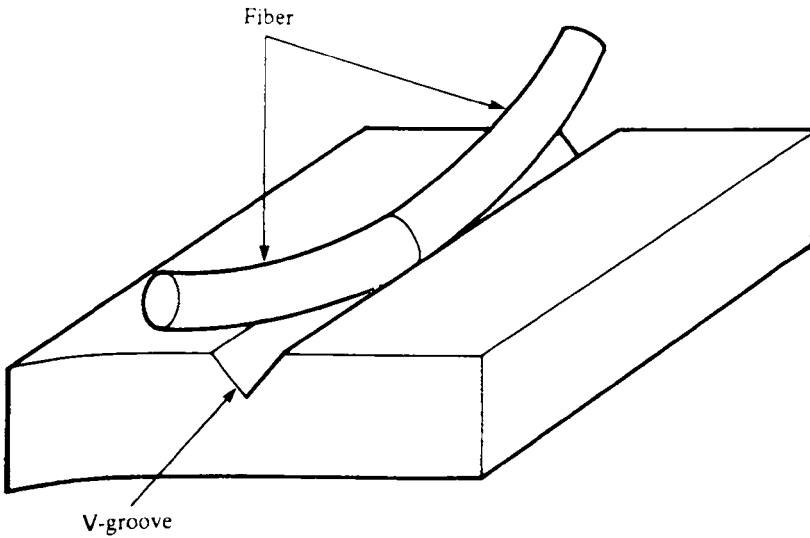


FIG. 8.36 Illustration of the use of a V-groove for the alignment of fiber ends prior to joining.

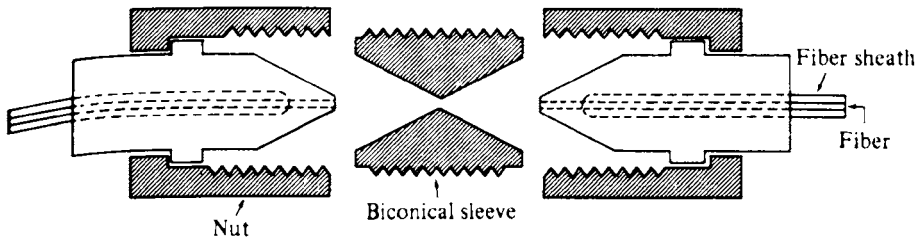


FIG. 8.37 Schematic diagram of a biconical demountable connector for butt joining optical fibers.

Typical values for loss in multimode connectors can range from a few decibels down to a few tenths of a decibel depending on the degree of precision in the alignment (and usually, in consequence, the cost). Losses in single mode connectors are generally somewhat higher than in their multimode counterparts – not surprisingly in view of their much smaller core diameters.

A type of demountable connector that does not rely on a butt join is shown in Fig. 8.38. This uses a beam expander in which each half of the connector contains a lens with the fiber end at the focal point. Light emerging from one end of a fiber becomes a relatively broad, collimated beam in the region between the lenses and is then focused back onto the other fiber end. Since the expanded beam width can be considerably larger than the fiber diameter, losses due to offset misalignments between the two halves of the connector are now considerably reduced. Great care is needed, however, in the initial alignment of the fiber with regard to the lens in each half of the connector. An advantage of this type of connector is that further optical components (e.g. beam splitters) can be inserted into the collimated beam between the lenses.

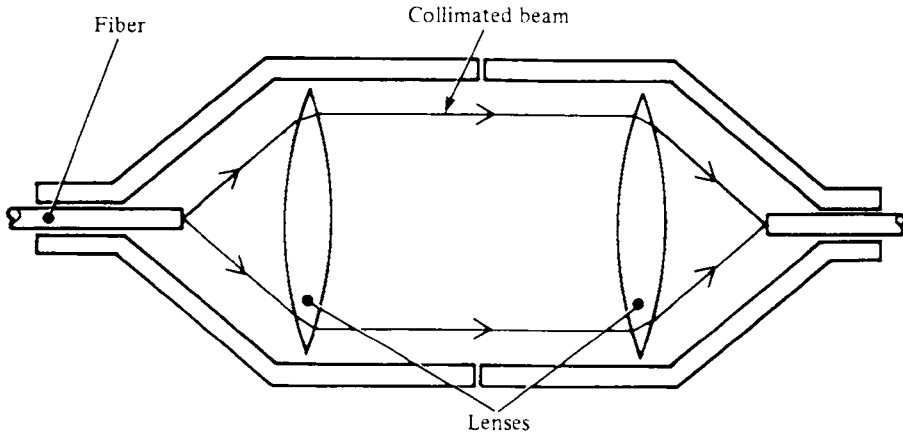


FIG. 8.38 Fiber coupling along a beam expander. The fiber ends are situated at the focal points of the two lenses resulting in an expanded collimated beam between the lenses.

We turn now to the problem of obtaining the requisite smooth, flat ends to the fiber. One of the simplest ways involves cleaving: the fiber is bent under tension over a curved surface whilst a knife edge ‘nicks’ the fiber (Fig. 8.39). This initiates crack formation in the fiber which, because of the curvature and tension in the fiber, rapidly spreads across the fiber in a direction perpendicular to the axis of the fiber. If carefully done end faces which are within a degree or so of the perpendicular to the fiber axis can be obtained. A more traditional technique is to polish the ends mechanically whilst they are held in a suitable jig. Again good accuracy can be obtained, but the process can be quite time consuming. Before either of these techniques is carried out the outer plastic coating on the fiber must be removed, either by using some type of mechanical stripper or by chemical means.

8.5.2 Fiber couplers

A problem arising in some applications of optical fibers is how to share a signal from one or more inputs amongst several output fibers. Often there are as many input fibers as there are output fibers. In this case the device is usually symmetrical and may be operated in ‘reverse’, with the output becoming the inputs and vice versa. A schematic illustration of a four-port device is shown in Fig. 8.40. Here, an input to either of fibers 1 or 2 will result in an output from fibers 3 and 4. The way in which the power from the input fiber is distributed amongst the output fibers is governed by the *insertion loss* L_{ins} , where L_{ins} is defined by

$$L_{ins} = -10 \log_{10}(P_j/P_i) \quad (8.41)$$

Here P_i is the power flowing in the i th input fiber and P_j is the power flowing in the j th output fiber. P_j will always be less than P_i for two reasons: first because of the splitting of the input power amongst the output fibers and secondly because there will inevitably be some

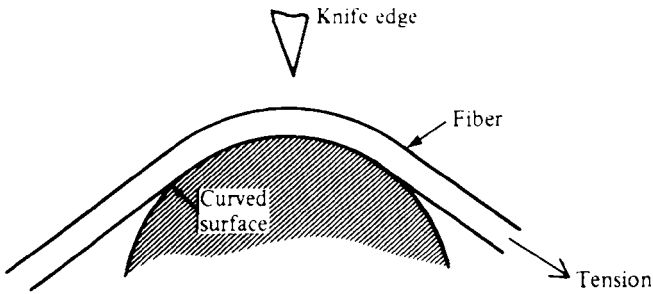


FIG. 8.39 Illustration of one technique used to prepare plane fiber ends normal to the fiber axis. The fiber is held in tension over a curved surface. Then the surface is scored with a tungsten carbide or diamond knife edge. This initiates crack formation and, provided the tension and radius of curvature are set correctly, a smooth break is obtained.

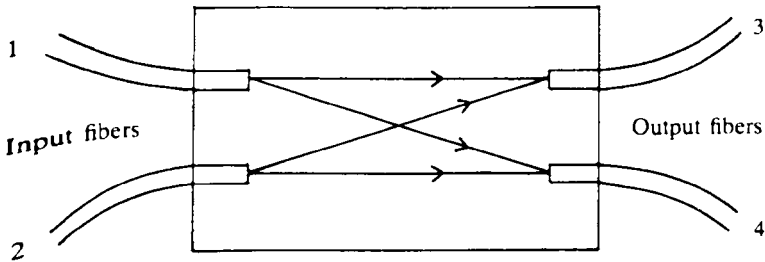


FIG. 8.40 Schematic representation of a four-port (2×2) fiber coupler. Power flowing in either fiber 1 or 2 will be distributed, in some predetermined way, into fibers 3 and 4.

loss of optical power within the device itself. The total power loss is described by the so-called *excess loss* (L_e) where

$$L_e = -10 \log_{10} \left(\frac{\sum P_i}{P_i} \right) \quad (8.42)$$

Another term used is the *output ratio*. This describes the nominal insertion loss from one input fiber to each of the output fibers (as in eq. 8.41) assuming there is no power loss in the device. It thus represents an ideal situation rather than the actual performance of the device.

A simple way of implementing such a coupler for multimode fibers is to butt the input fibers up against the end of a 'mixing rod'. This is essentially a short length of large diameter optical fiber. The output fibers are similarly butted up against the other end of the rod (Fig. 8.41a). The rod must be long enough to ensure that the output from each input fiber becomes sufficiently diffused within the rod to give an even spread of energy over the output face. To avoid excessive loss, it is best if the fibers can be 'close packed' over the end of the rod, a situation that can only be achieved for certain numbers (e.g. 7, 19, etc.) of input fibers (see Fig. 8.41b). A calculation of the losses resulting from this type of coupler is given in Example 8.14.

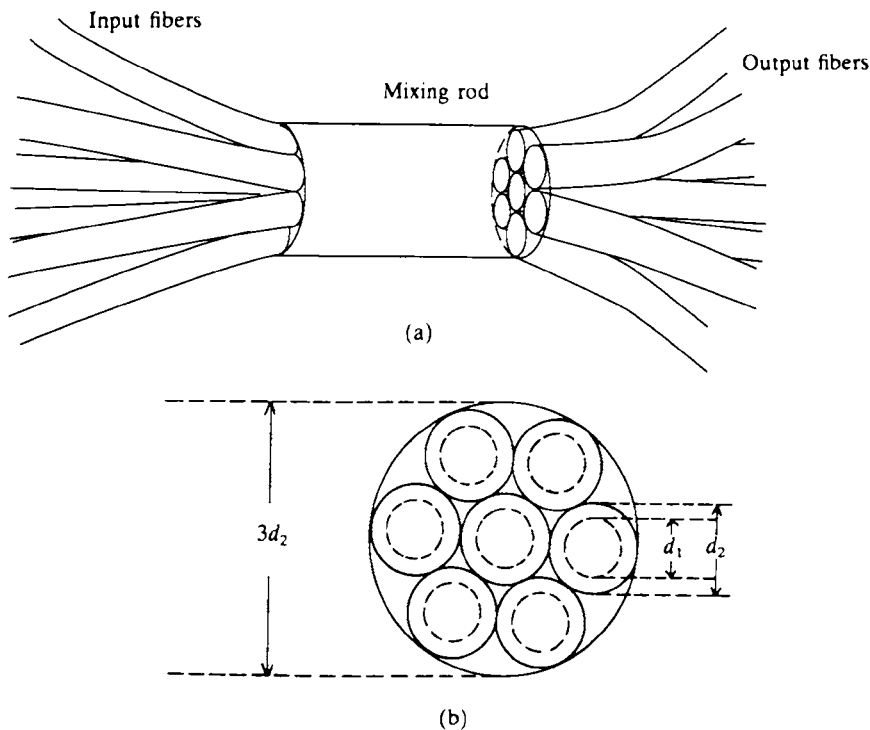


FIG. 8.41 Design for a 14 port (7×7) fiber coupler using a 'mixing rod'. The overall scheme is shown in (a), whilst (b) shows how seven fibers can form a 'close-packed' array across the face of the mixing rod.

EXAMPLE 8.14 Fiber coupler losses

We consider the rod-type coupler illustrated in Fig. 8.41(a) with seven input and output fibers, each having core and core + cladding diameters of d_1 and d_2 respectively. If the mixing rod has a diameter of $3d_2$, the close-packed arrangement of Fig. 8.41(b) can be obtained. Losses will inevitably result because the cores of the output fibers have a combined area that is less than that of the end of the mixing rod. Assuming a uniform spread of light from any one input fiber over the end of the coupling rod and no losses in the coupling rod, we may write $P_i = B\pi(3d_2)^2/4$ whilst $P_o = B\pi d_1^2/4$, where B is a constant. Thus from eq. (8.41) the insertion loss is given by

$$L_{\text{ins}} = -10 \log_{10} \left(\frac{B\pi d_1^2/4}{B\pi(3d_2)^2/4} \right) \quad \text{or} \quad -10 \log_{10} \left(\frac{d_1^2}{9d_2^2} \right)$$

Taking $d_1 = 200 \mu\text{m}$ and $d_2 = 250 \mu\text{m}$ gives $L_{\text{ins}} = 11.5 \text{ dB}$.

Similarly the excess loss is readily seen to be given by

$$L_e = -10 \log_{10} \left(\frac{7d_1^2}{9d_2^2} \right)$$

whence $L_e = 3 \text{ dB}$.

Another technique used to make multimode fiber couplers is to twist the fibers around each other and then to heat them under tension in an oxypropane torch flame. The bundle fuses together and then elongates to form a device known as a *fused biconical taper coupler*. The coupling mechanism is believed to be as follows: when light from one fiber enters the fused region, the narrowing core causes mode conversion into cladding modes. These subsequently spread out across the whole of the cross-section in the fused region. Then, when the fibers separate, the cladding modes are converted back into core modes, but with the energy now distributed amongst all the fibers. Such couplers exhibit low coupling losses but the coupling properties between the various ports may vary.

Single mode fiber couplers can also be made by the same technique, though the coupling mechanism is thought to be somewhat different. When the radiation encounters the narrowing fiber core, the effective V parameter of the fiber will decrease (since $V \propto d/\lambda_0$) and the mode field will extend further out into the cladding (e.g. see Fig. 8.25) so that it overlaps with the other fiber cores, and hence causes energy to be coupled into them. This is illustrated in Fig. 8.42.

By using the so-called *coupled mode theory* (ref. 8.28) it may be shown that the fraction, F , of the energy initially in one of the fibers which is coupled into the other can be written

$$F = \sin^2(Cz) \quad (8.43)$$

where z is the distance over which the fibers are coupled and C represents the strength of the coupling between the fibers. The latter should depend on the spread of the evanescent field which in turn depends on the V parameter: the smaller the value of V the greater the spread. Since V is inversely proportional to the wavelength we may conclude that the longer the wavelength the stronger the coupling. Thus C will be some increasing function of wavelength, $C(\lambda)$, which we assume we can write as $b\lambda^q$, where b and q are constants, so that

$$F = \sin^2(b\lambda^q z)$$

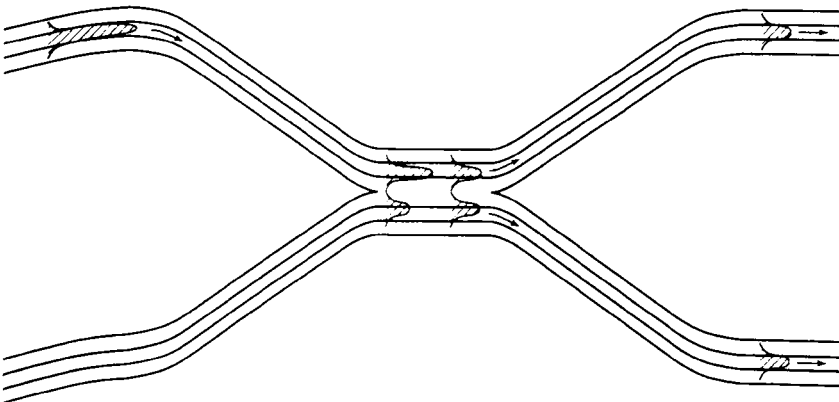


FIG. 8.42 Coupling mechanism in a single mode biconical taper coupler involving evanescent mode overlap.

The coupler will exhibit a quasi-oscillatory behaviour as a function of wavelength, thus emphasizing that an intended coupling ratio will only be obtained at a particular operating wavelength.

One of the most useful single mode couplers is a four-port device where a given input signal is split equally between two output fibers. Assuming no losses in the coupler, the insertion loss is 0.5 or 3 dB. Such couplers are referred to as 3 dB couplers.

8.6

Measurement of fiber characteristics

8.6.1 Introduction

In this section, we describe techniques used to measure those properties such as loss and bandwidth which are important in optical fiber communications. Measurements of purely mechanical properties are not included here for reasons of space (ref. 8.18 may be consulted for a fuller discussion of all types of measurement). Before addressing ourselves to specific techniques, however, it is useful to understand the importance of obtaining the correct launch conditions.

Suppose we illuminate the whole of a fiber end with a uniform, highly convergent beam of light. All modes in the fiber will then be excited, including leaky and cladding modes. If we were to examine the mode structure a considerable distance along the fiber, almost all of the lossy modes would have died out. In addition, mode coupling would have thoroughly 'scrambled' the remaining modes to generate an 'equilibrium' mode distribution that would then remain more or less the same throughout the rest of the length of the fiber. In principle, such a distribution would be obtained whatever the launch conditions, provided we make our observation of the mode distribution sufficiently far along the fiber. In practical terms, however, this might amount to several kilometres. Most of the properties we are concerned with here do indeed depend of the mode distribution, so that it makes sense to try and measure these for the equilibrium distribution. On the other hand, it may be very inconvenient if many kilometres of fiber have to be used to obtain this distribution. It would obviously be preferable if the launch conditions were such as to set up equilibrium right from the start. Unfortunately this is not at all easy to do, although several procedures can be adopted that go some way towards realizing this goal. For example, to avoid launching cladding and leaky modes care must be taken to illuminate only the core area with radiation whose angle of incidence is less than the acceptance angle of the fiber (as given by eq. 8.21).

A device that aims to produce an equilibrium 'mix' of modes over a short length of fiber is the *mode scrambler*. In this, a large amount of mode coupling is deliberately introduced by means of bending or microbending distortion (see section 8.4.1). For example, a few turns of the fiber may be wrapped round a post or mandrel a few tens of millimetres in diameter (Fig. 8.43a). Alternatively, the fiber may be passed through a series of posts to produce a serpentine distortion (Fig. 8.43b). A simple variant of the latter technique is to sandwich the fiber between two layers of coarse sandpaper over a length of some tens of millimetres.

The use of a mode scrambler may itself cause cladding modes to be excited, and these will have to be removed subsequently. One device designed to remove cladding modes is

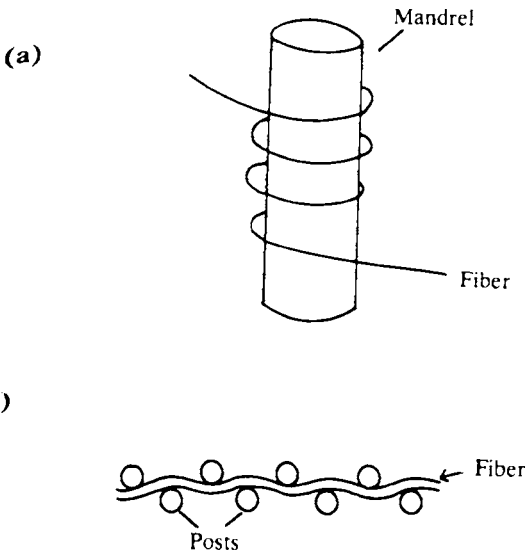


FIG. 8.43 Techniques for producing an equilibrium mode distribution. In (a), the 'mandrel wrap' method, a few turns of fiber are taken around a rod. In (b), the fiber makes a serpentine path through a series of posts.

called a *mode stripper*, which may be simply constructed by taking a short length of fiber that has been stripped of its outer protective coating and pressing it against a felt pad that has been saturated in oil (such as paraffin oil) which has a slightly higher refractive index than the cladding. Any light incident on the cladding/oil interface will then be refracted out of the fiber and absorbed within the felt pad.

8.6.2 Fiber attenuation measurements

At its simplest, the measurement of attenuation involves focusing a known amount of monochromatic radiation onto the end of a known length of fiber and then measuring the amount emerging from the other end (Fig. 8.44). Use of eq. (8.37) then enables the fiber attenuation to be determined. Radiation at different wavelengths can be obtained by using a white light source, such as a quartz halogen lamp, in conjunction with either a monochromator or a set of narrow bandpass optical filters. This method ignores the inevitable losses incurred when coupling light into the fiber, although for fibers with high attenuation, this may prove acceptable. With low loss fibers, however, the *cutback* technique is generally used. In this, light is launched as in the simple technique and the emerging power measured. Then, without changing the launch conditions in any way, the fiber is cut back to within a few metres of the launch end, and the emerging energy again measured. The fractional transmittance of the length of fiber removed is then given by the ratio of these two measurements, since the launch energy is the same in both cases. As outlined in the introduction to this section, care must be taken with the launch conditions to obtain an equilibrium mode distribution (i.e. by using a mode scrambler/mode stripper combination).

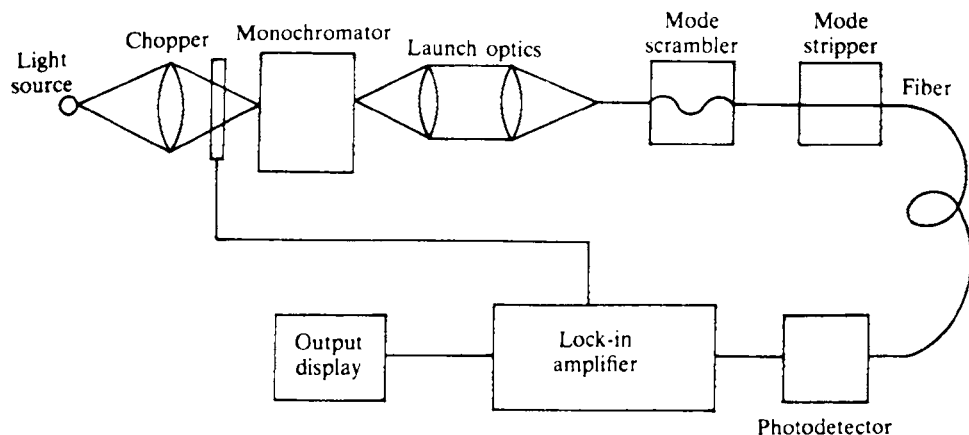


FIG. 8.44 Schematic layout of apparatus for measuring attenuation in optical fibers. To enhance the signal-to-noise ratio, an optical chopper–lock-in amplifier combination is often used.

The above techniques measure the total attenuation within the fiber, which, from a practical point of view, is usually all that is required. However, it may be that the individual contributions from absorption and scattering need to be assessed separately. Absorption losses may be determined by measuring the temperature rise within a given length of fiber, whilst scattering loss measurements depend on determining the amount of light scattered out of the fiber (see Problem 8.18 and also ref. 8.19).

8.6.3 Fiber dispersion measurements

When a narrow optical pulse is launched into a fiber, the various dispersion mechanisms discussed in sections 8.3.2 to 8.3.5 will cause the pulse to broaden as it travels down the fiber. In the so-called *time domain* measurement of dispersion, this broadening is measured directly. A narrow optical pulse may conveniently be generated by using a semiconductor laser in conjunction with fast driving electronics. Pulses with widths of a few nanoseconds or less may be readily obtained. If the main source of dispersion is intermodal dispersion then, as in attenuation measurements, an equilibrium mode distribution will have to be created. The detector used must obviously have a fast response and often avalanche photodiodes are used. Since the pulses may be too narrow to be displayed easily on an ordinary oscilloscope, a fast sampling oscilloscope may be required.

Ideally, the input pulse should be narrow compared with the output pulse so that it can be thought of as a 'spike' of negligible width. The shape of the output pulse is then called the *transfer function* (see Appendix 5). Normally, however, the input pulse has appreciable width and determining the transfer function becomes more difficult. A simplification often adopted is to assume that the input pulse, transfer function and output pulses are all Gaussian in shape (see Fig. A5.1b); that is, we can describe each curve by the function $A(t) = A(0) \exp(-t^2/2\sigma^2)$, where time is measured from the peak of the curve. The parameter σ essentially determines the r.m.s. pulse width. If σ_i , σ_t and σ_o are the values for the initial,

transfer and final pulse widths, then it can be shown (ref. 8.20) that

$$\sigma_f^2 = \sigma_i^2 + \sigma_t^2 \quad (8.44)$$

that is,

$$\sigma_t = (\sigma_f^2 - \sigma_i^2)^{1/2} \quad (8.44a)$$

An alternative method of measuring dispersion is to take measurements in the frequency domain. An emitting device (a laser or LED, for example) is amplitude modulated with a variable frequency sinusoidal waveform, and the output from the fiber is recorded as a function of frequency. This allows the frequency bandwidth of the fiber to be measured directly. In fact two bandwidths are commonly used: the optical bandwidth Δf_{opt} , and the electrical bandwidth Δf_{el} . The former is defined as the frequency at which the signal falls to half its maximum low frequency value, the latter as the frequency when the signal falls to $1/\sqrt{2}$ of its maximum value. Appendix 5 should be consulted for a discussion of the relationships between pulse dispersion and bandwidths. For a Gaussian pulse, the electrical bandwidth and pulse width are related by

$$\Delta f_{\text{el}} = \frac{1}{7.56\sigma}$$

Although the two techniques above enable the total dispersion exhibited by a fiber to be determined it is evident that since material dispersion is such an important pulse broadening mechanism that it deserves a more detailed investigation. The most usual technique is to measure the time taken for a pulse to traverse a known length of fiber as a function of the wavelength of the source generating the pulse (this assumes we have a variable wavelength source, or a series of sources of differing wavelengths). This then determines how the group transit time (τ_g) depends on λ_0 . The gradient of this relationship then gives the quantity $d\tau_g/d\lambda_0$ as a function of λ_0 . This, of course, is the quantity which is required in eq. (8.34) to determine pulse spreading due to material + waveguide dispersion.

8.6.4 Cut-off wavelengths in single mode fiber

The cut-off wavelength, λ_c , is the wavelength in a fiber below which it ceases to behave as a single mode fiber. One of the simplest (but not necessarily the most accurate) techniques to measure λ_c is to launch light of variable wavelength obtained from, say, a broad band radiation source and a monochromator into the fiber, and measure the transmission of the fiber as a function of wavelength. With two modes travelling in the fiber the amount of radiation that can be transported along the fiber is roughly doubled compared with when only a single mode can propagate. Thus the transmission curve shows a sudden increase below the cut-off wavelength λ_c , as shown, for example, in Fig. 8.45. Allowance should be made for the system (i.e. source/monochromator/detector) response. Care must also be taken to ensure that the fiber is not unduly bent, since the higher order modes show heavy bending losses when they are near cut-off, and may not be able to propagate at all if the fiber is bent too much. In fact a related method of determining λ_c is to compare the transmission of the same length of fiber between when it is straight and when it is bent. This technique obviates the need to allow for system response.

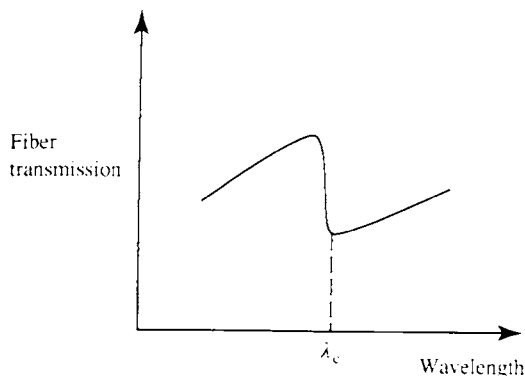


FIG. 8.45 Diagram of the relative amount of power transmitted through a single mode fiber as a function of wavelength. There is a sudden change in the transmission at the cut-off wavelength, λ_c .

8.6.5 Refractive index profile measurement

Several methods are available for measuring the fiber core refractive index profile, none of which, however, is entirely satisfactory. Probably the most accurate method is the interferometric slab technique. In this a thin slice of fiber of thickness d is taken perpendicular to the fiber axis. The thickness must be uniform to within a fraction of the wavelength of the light used. Light passing through the slab will normally 'see' an optical thickness of $n(r)d$, where $n(r)$ is the refractive index a radial distance r from the fiber center. The slab is then placed in the arm of an interferometric system (see section 6.6.1). The non-uniform optical path differences through the slab then show up as distortions in the uniform fringe pattern observed in the absence of the slab. Using this technique, the refractive index can be measured to better than one part in 10^4 , the main limitation being the accuracy with which the thickness of the slab can be measured. The disadvantages are that it requires precise and time-consuming sample preparation and it is also, of course, destructive to the fiber.

A non-destructive technique capable of yielding reasonably accurate results is the near-field scanning method. If all the guided modes are uniformly excited then the variation of light irradiance across the end of the fiber may be written (ref. 8.21) as

$$I(r) = I(0) \left(\frac{n_1^2(r) - n_2^2}{n_1^2(0) - n_2^2} \right) \quad (8.45)$$

Here $I(r)$ is the irradiance observed a distance r from the center, and $n_1(r)$ and n_2 are the core and cladding refractive indices respectively. If any leaky modes are present, however, a correction factor has to be applied (ref. 8.22). The apparatus requirements are reasonably straightforward. One end of the fiber is illuminated by focusing a Lambertian source (e.g. a tungsten filament lamp) onto it. This ensures that the guided modes are all excited equally. A magnified image of the other end of the fiber is formed in the plane of a small area photodiode, which is then scanned across the image allowing $I(r)$ to be determined.

The two methods described here are by no means the only ones that can be used to determine refractive index profiles; other techniques will be found in ref. 8.23.

8.6.6 Optical time domain reflectometer

All the measurement techniques described hitherto are essentially laboratory based and would be difficult to implement in the field. Furthermore, they often require access to both ends of the fiber, a situation causing not a little difficulty if the fiber ends are several kilometres apart! One instrument capable of yielding considerable information about an optical fiber system, and yet which requires access to one end only of the fiber, is the optical time domain reflectometer (OTDR).

The basis of this instrument is that a narrow pulse of radiation is launched into one end of the fiber, and then the amount of radiation re-emerging from this same end is monitored as a function of time after the initial launch. As the pulse travels down the fiber, it will suffer attenuation so that after a length L of fiber the pulse energy will be given by $E(L)$, where

$$E(L) = E(0) \exp(-\alpha L)$$

where α is the absorption coefficient. At every point along the fiber, a certain portion, S_R say, of the pulse energy will be lost by Rayleigh scattering. Of this a fraction, f_g (see Problem 8.21), will enter guided modes propagating back down the fiber. The amount of energy scattered back down the fiber may therefore be written $S_R f_g E(L)$. This backscattered radiation will itself be subject to attenuation as it travels back to the launch end of the fiber. The energy finally re-emerging owing to backscattering from a point a distance L into the fiber is given by $E_b(L)$, where

$$E_b(L) = S_R f_g E(L) \exp(-\alpha L)$$

or

$$E_b(L) = S_R f_g E(0) \exp(-2\alpha L)$$

Thus

$$\log_e E_b(L) = \log_e [S_R f_g E(0)] - 2\alpha L \quad (8.46)$$

Now the time delay, t , between the initial launch of the pulse and the subsequent return of the backscattered energy from L may be approximated by $t = 2Ln_1/c$, so that we may transform eq. (8.46) to give

$$\log_e E_b(L) = \log_e [S_R f_g E(0)] - \frac{\alpha t}{n_1} \quad (8.47)$$

Thus a plot of the logarithm of the returning energy as a function of time should yield a straight line with gradient $-\alpha/n_1$. If n_1 is known, α may be determined.

In this analysis, we have assumed a perfectly uniform fiber as far as α , S_R and f_g are concerned. In practice, α may well vary along the fiber and discontinuities, such as fiber joins, will give local increases in the amount of backscattering. For example, a butt join will give something like a 4% backscatter because of Fresnel reflection. These will show up on the graph as 'spikes'. In addition, the power loss within a join or splice will lead to a corresponding reduction in the signal backscattered from beyond it. The end of the fiber will be marked by a Fresnel spike with the signal falling to zero immediately afterwards. If α does

in fact vary along the length of the fiber, the above graph will not be a straight line; however, the local gradient should reflect the local value of α . In commercial instruments, the returning energy is measured in decibels and time is converted into distance along the fiber. Figure 8.46 shows a typical trace.

One of the main difficulties in realizing this instrument is that the amount of energy returning from the fiber is very small, and even under optimum conditions it is difficult to examine more than 15 km of fiber. The light source is invariably a pulsed semiconductor laser, whilst the returning radiation is measured with an avalanche photodiode. Both to launch light and to measure the returning light, some form of beam splitter arrangement is required. Usually, the instrument sends out a continuous stream of pulses. After each of these, the amount of energy returning after a certain (variable) time delay is measured. The time delay is gradually increased, thus enabling the whole curve to be built up point by point. To increase accuracy, several readings may be taken for each delay and an average taken. Typical pulse widths are the order of 50 ns with launched energies of a few hundred milliwatts per pulse. Features on the curve can be located to within ± 5 m, or perhaps a little better. Overall, a dynamic range of something like 40 dB ($= 2 \times 20$ dB single way loss) can be tolerated over the length of fiber under inspection before the returning signal becomes too weak to measure.

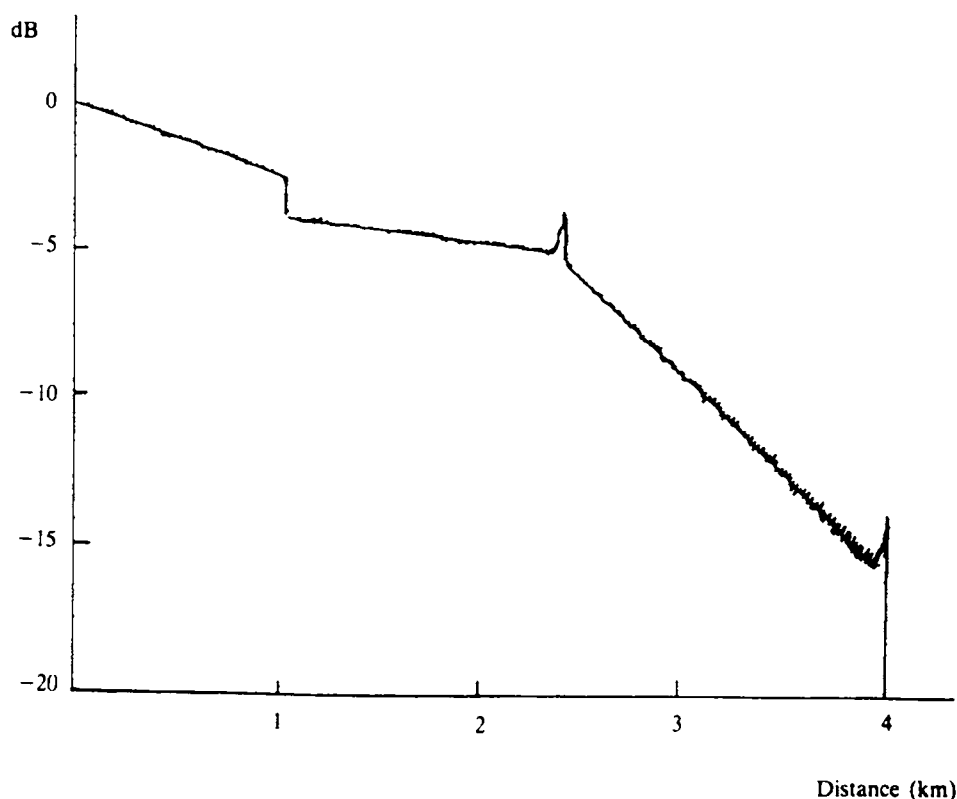


FIG. 8.46 Typical trace from an OTDR.

8.7

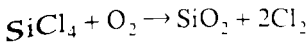
Fiber materials and manufacture

Only two main materials have been seriously considered to date for use in optical waveguides, these being various types of plastics and silica. Plastic fibers offer some advantages in terms of cost and ease of manufacture, but their high transmission losses preclude their use in anything other than short distance (i.e. less than a few hundred meters) optical links.

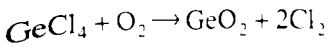
8.7.1 Silica-based fibers

An obvious requirement in any type of fiber manufacture is that the core refractive index must be higher than that of the cladding. Pure silica has a refractive index of about 1.45 at $1\ \mu\text{m}$ and this may be changed by doping with various impurities. For example, impurities that raise the refractive index are titania (TiO_2), alumina (Al_2O_3) and germania (GeO_2), whilst boria (B_2O_3) and fluorine (F) lower it. The solubility of such materials in silica is limited and the maximum change in refractive index is limited to a few per cent. Some of the highest performance fibers have been fabricated with germania-doped cores and fluorine-doped claddings.

Most techniques currently used to manufacture silica fiber are based on some type of vapour deposition technique. One of the most popular of these is the so-called *modified chemical vapour deposition* (MCVD) method. In this a doped silica layer is deposited onto the inner surface of a pure silica tube. The deposition occurs as a result of a chemical reaction taking place between the vapour constituents that are being passed down the tube. Typical vapours used are SiCl_4 , GeCl_4 and O_2 and the reactions that take place may be written



and



The zone where the reaction takes place is moved along the tube by locally heating the tube to a temperature in the range $1200\text{--}1600^\circ\text{C}$ with a traversing oxyhydrogen flame (Fig. 8.47). If the process is repeated with different input concentrations of the dopant vapours, layers of different impurity concentrations may be built up sequentially. Once the deposition process is complete, any residual gases are pumped out and the tube is heated to its softening temperature ($\approx 2000^\circ\text{C}$). Surface tension effects cause the tube to collapse into a solid rod called a *preform*.

A fiber may subsequently be produced by drawing from the heated tip of the preform as it is lowered into a furnace (Fig. 8.48). To exercise tight control over the fiber diameter, a thickness monitoring gauge is used before the fiber is drawn onto the take-up drum, and feedback is applied to the drum take-up speed. In addition, a protective plastic coating is often applied to the outside of the fiber (see Fig. 8.48 and section 8.8) by passing it through a bath of the plastic material; the resulting coating is then cured by passing it through a further furnace.

In other vapour deposition techniques, the preform may be made by growing doped silica layers on the outside of a silica rod. The reactant gases, such as SiCl_4 and GeCl_4 , are fed

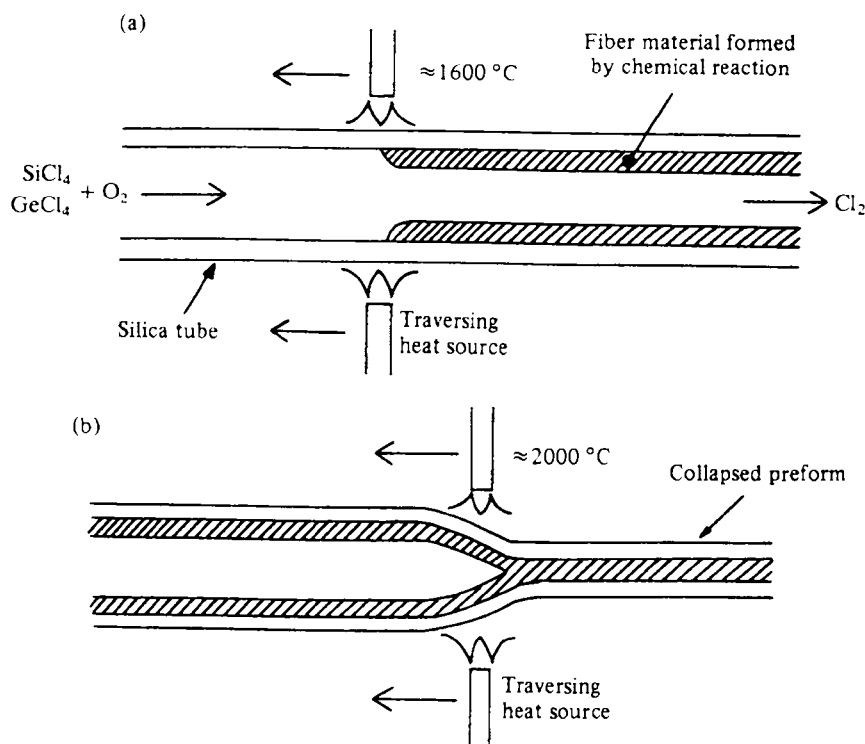


FIG. 8.47 Production of fiber preform by modified chemical vapour deposition. In the first stage (a) the reactants are introduced into one end of a silica tube and the core material deposited on the inside of the tube in the reaction zone, where the temperature is maintained at about 1600°C . Several traverses of the heating assembly may be necessary to build up sufficient thickness of core material. In the second stage (b) the tube is collapsed into a solid preform rod by heating to the silica softening temperature (about 2000°C).

into H_2/O_2 flames where a stream of fine doped silica soot particles is formed. In the *outside vapour deposition* (OVD) technique, these are deposited on the outside of a rotating mandrel, the growth taking place radially. Alternatively, as in the *vapour axial deposition* (VAD) technique, deposition may take place on the end surface of a silica 'seed' rod. Both of these deposition techniques produce a porous material, and so a further production stage is required involving heating the rod to between 1000°C and 1500°C in the presence of chlorine. The solid preform thus formed may be drawn into a fiber in exactly the same way as described above for the MCVD technique.

8.7.2 Plastic-coated silica fiber

Fibers are available which have a silica core combined with a plastic cladding. These *plastic-coated silica* (PCS) fibers are somewhat easier to manufacture than all-silica fiber. A preform is made consisting of the core material only. The preform to fiber process then proceeds as above, the difference now being that the plastic bath provides the cladding material rather

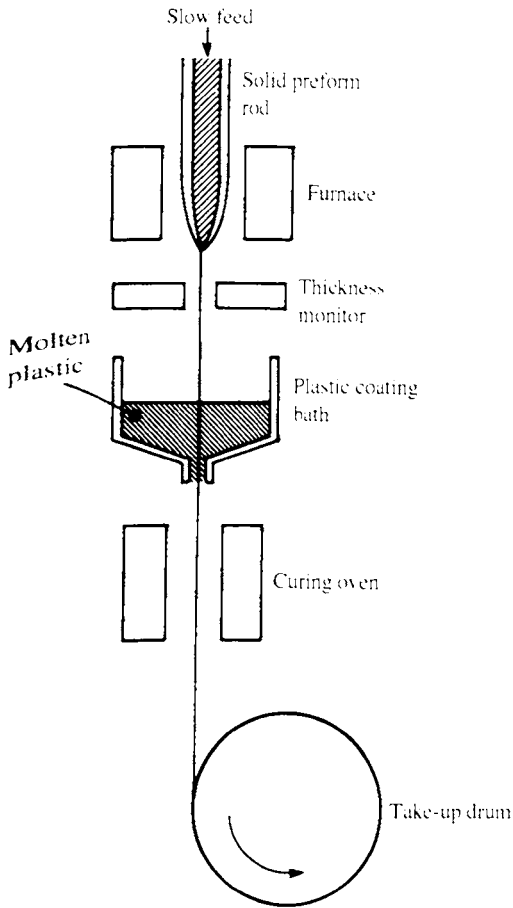


FIG. 8.48 Conversion of a preform rod into a fiber. The end of the rod is heated to its softening point and the fiber drawn off from the tip and wound onto a take-up drum. The fiber is usually coated with a layer of plastic for protection before being wound on the drum.

than a protective coating. The process readily lends itself to the production of step index multimode fibers with relatively large core diameters where little of the energy is carried in the cladding. Such fibers are attractive for medium distance, moderate bandwidth communication systems, where cost is a major consideration. Typical losses are of the order of 10 dB km^{-1} . These values are significantly higher than those in all-silica fibers because, although very little of the energy travels in the cladding, that which does is subject to much higher attenuation.

8.7.3 All-plastic fibers

Fibers can also be made entirely from plastic materials, a popular combination being a core of polymethyl methacrylate (PMMA, $n = 1.495$) and a cladding of fluoroalkyl

methacrylate ($n = 1.402$). Compared with silica fibers losses in such fiber are high (Fig. 8.49) but the basic causes are the same: there are strong absorption bands associated with vibrations of the C–H bond and also high Rayleigh scattering from the long chain molecules. It will be seen from Fig. 8.49 that transmission windows exist at 570 nm and 650 nm. Although attenuation is in fact lower for the former wavelength, plastic fibers are almost always used in conjunction with red-emitting LEDs. This reflects the fact that these have a higher output and a superior frequency response than green LEDs. The ultimate attenuation in plastic fibers is not known with any great certainty, although values as low as 10 dB km^{-1} have been estimated. Considerable effort is being put into lowering the attenuation in these fibers (ref. 8.24).

In spite of their (current) high attenuation all-plastic fibers do offer some significant advantages when used over distances of a few hundred metres or so. Because of their increased flexibility they can be made with diameters considerably greater than silica fibers (up to 3 mm, although 1 mm is more normal). This makes coupling fibers together much easier, and simple injection moulded connectors can be used. Fiber end preparation is also very straightforward: simple cutting with a sharp blade is usually adequate. In addition the relatively large differences between core and cladding refractive indices give rise to high numerical apertures (typically 0.5), which makes it easy to couple light in from sources such as LEDs. Techniques have also been developed for making graded index plastic fibers which offer the possibility of very high bandwidths over short distances, but these are not, at the time of writing, available commercially.

One of the problems with PMMA-based fiber is that it can only be operated in environments below about 70°C . This precludes its use in, for example, the automobile engine for monitoring purposes etc. Other alternatives, for example polycarbonates, are being investigated, although these usually exhibit higher attenuation.

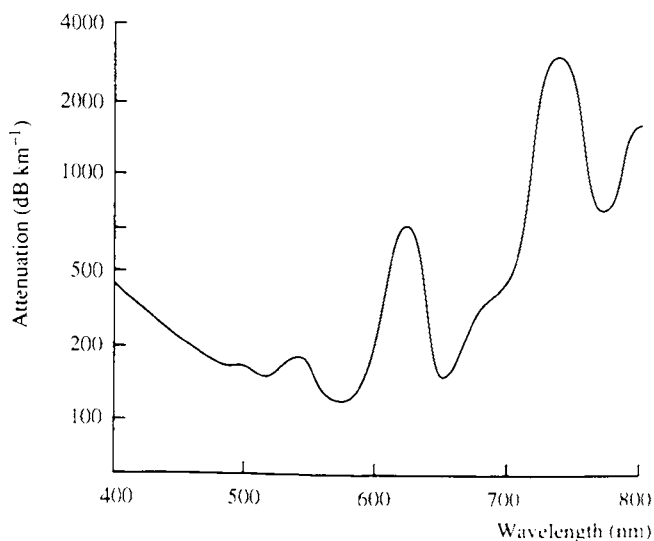


FIG. 8.49 Typical attenuation versus wavelength plot for an all-plastic fiber.

8.7.4 Mid-infrared fibers

The performance of silica-based fibers has been dramatically improved over the last few years to the stage where, over a wide wavelength range, the theoretical attenuation limit given by Rayleigh scattering has, to all intents and purposes, been reached. As we have seen the lowest attenuation (about 0.2 dB km^{-1}) occurs at a wavelength of $1.55 \text{ }\mu\text{m}$. This is as a result of the presence of Rayleigh scattering, which declines with increasing wavelength (as λ^{-4}), and vibronic absorption, which produces a sudden rapid increase beyond a particular wavelength. If materials could be found where the vibronic absorption peaks occurs further into the infrared and where Rayleigh scattering is at least as low as in silica, then there would be the possibility of obtaining fibers with even lower minimum attenuations than are possible at present. For example, if the vibronic absorption in silica set in at $3.2 \text{ }\mu\text{m}$ rather than at $1.6 \text{ }\mu\text{m}$, then the minimum attenuation should be a factor 2^4 or 16 smaller than in silica, that is about $0.2/16$ or $0.0125 \text{ dB km}^{-1}$. Figure 8.50 shows the theoretical attenuations for a number of promising materials.

Interest has centred mainly round two types of material, namely the heavy metal fluoride systems and the chalcogenide glasses (ref. 8.25). One of the most studied examples of the former is $\text{ZrF}_4\text{-BaF}_2\text{-LaF}_3\text{-AlF}_3\text{-NaF}$ or 'ZBLAN'. So far minimum attenuations of between 1 dB km^{-1} and 2 dB km^{-1} have been obtained at a wavelength of $2.5 \text{ }\mu\text{m}$. These are about two

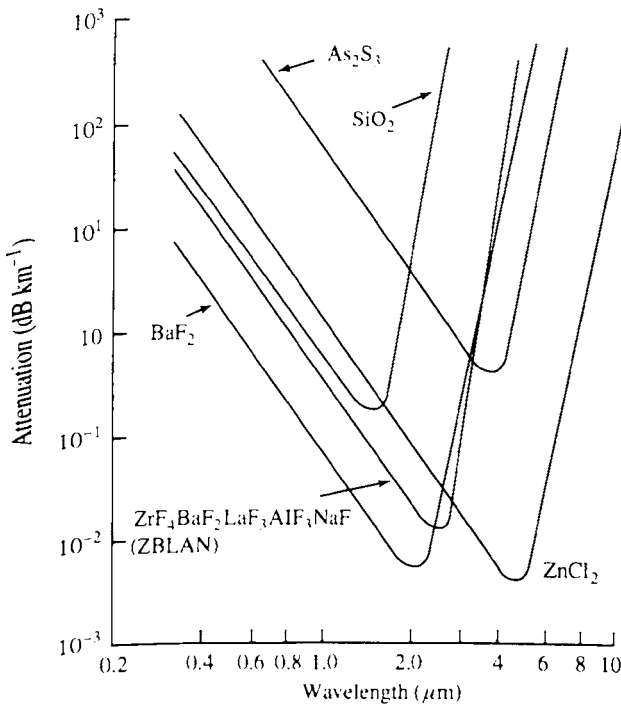


FIG. 8.50 Predicted loss as a function of wavelength for some of the candidates for use in low loss mid-infrared fibers.

orders of magnitude larger than the theoretical minimum; the main problem would seem to be scattering from submicrometre scattering centres such as clumps of impurities and bubbles.

Chalcogenide glasses consist of compounds formed between one of the elements Ge, Si, As, Sb and one of the elements S, Se, Te, an example being As_2S_3 . These offer the prospect of low attenuation at yet higher wavelengths (up to, say, $10\text{ }\mu\text{m}$). Here again the actual transmission losses obtained in practice are orders of magnitude higher than the theoretical values. In spite of a sustained worldwide research effort, it has not been possible so far to find any better material for optical fibers than silica. Current opinion would seem to be that as far as long distance communication links are concerned we will have to make do with silica for many more years to come. It should also be pointed out that if such research is successful, then new detector and emitter materials would have to be developed.

Not all uses of fiber, however, require long lengths; there are many instances where transmission of mid-infrared radiation over just a few metres is all that is needed. One example of this is the use of the CO_2 laser in surgery where the availability of a flexible beam delivery system such as the optical fiber is extremely useful (see section 6.8.2 and ref. 8.26).

8.7.5 Special fiber types

8.7.5.1 Polarization-maintaining fiber

Although we frequently talk of single mode fibers, in fact *two* orthogonally polarized modes can usually propagate equally well down ordinary single mode fiber. In a circularly symmetrical fiber, these two modes will have identical propagation velocities. However, real fibers will not be perfectly symmetrical; there are bound to be slight anisotropies in both shape and refractive index. The result is that the two modes will inevitably have slightly differing velocities. In other words, the fiber is to some extent birefringent (section 3.2). The birefringence may result either from the intrinsic properties of the fiber (intrinsic birefringence) or from external perturbations such as bending or squeezing (extrinsic birefringence).

Let us suppose that we launch equal amounts of energy into the two orthogonally polarized modes of a single mode fiber. As the modes travel along the fiber their phase relationship will change (Fig. 8.51), but after a certain length (called the *beat length*) the phase difference reaches 2π and the original input polarization state recurs.

Suppose the velocities of the two modes can be written as c/n_x and c/n_y where n_x and n_y are the effective refractive indices for the two modes. After a distance L , the phase difference between the modes $\Delta\phi_p$ is given by

$$\Delta\phi_p = (n_x - n_y) \frac{2\pi L}{\lambda_0}$$

The beat length L_p is then given by the condition that $\Delta\phi_p = 2\pi$; hence

$$L_p = \frac{\lambda_0}{(n_x - n_y)} \quad (8.48)$$

Typical 'circular'-cored single mode fibers have beat lengths in the range $100\text{ mm} < L_p < 5\text{ m}$.

Now, modes that have very similar propagation constants can be coupled together by fiber

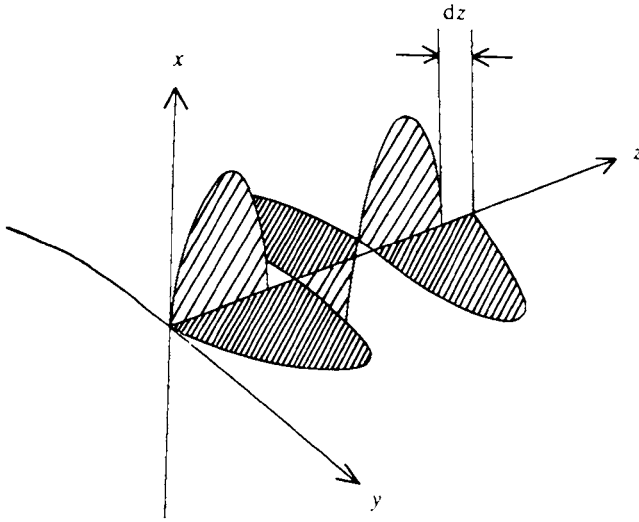


FIG. 8.51 Fiber birefringence causes a phase shift ($\Delta\phi = (2\pi/\lambda)dz$) to develop between the mode polarized along the y direction (heavy shading) and that polarized along the x direction (light shading) as the modes travel along the fiber in the z direction.

perturbations which have relatively long periodicities (see eq. 8.36 and subsequent discussion). Such perturbations are difficult to avoid and hence coupling between the two orthogonally polarized modes becomes a likely possibility. Thus, if one polarization state is launched into one end of a fiber, mode coupling will ensure that some of this mode energy will be coupled into the orthogonal mode. However, if the fiber has a large birefringence, the two polarization modes can be coupled only by very short period perturbations, which are much less likely to be present. Such fibers are known as 'Hi-Bi' fibers and are capable of maintaining the initial linear polarization state over quite large distances.

The quality of a polarization-maintaining fiber may be judged by the value of the polarization-holding parameter, h , which is defined as the amount of polarization mode coupling per unit length. Thus if $P_x(z)$ and $P_y(z)$ represent the two mode powers at a point z , then

$$\frac{dP_x(z)}{dz} = hP_y(z)$$

Assuming that h is small, $P_y(z)$ will change little with z , and so $P_x(z) = hzP_y(z)$. Fibers have been fabricated with h values as small as $5 \times 10^{-5} \text{ m}^{-1}$. This implies, for example, that after 1 km and assuming $P_x(0) = 0$, $P_x/P_y = 10^3 \times 5 \times 10^{-5}$, which is 5×10^{-2} or -13 dB in optical power terms.

Most polarization-maintaining fibers are intrinsically birefringent and several different techniques have been used to produce them. All rely on producing a fiber cross-section which is in some way asymmetric about two perpendicular axes. Usually, this asymmetry is built into the fiber preform before the fiber is drawn. Some examples are shown in Fig. 8.52. Elliptically cored fibers (Fig. 8.52a) have been made with beat lengths of about 1 m, but they

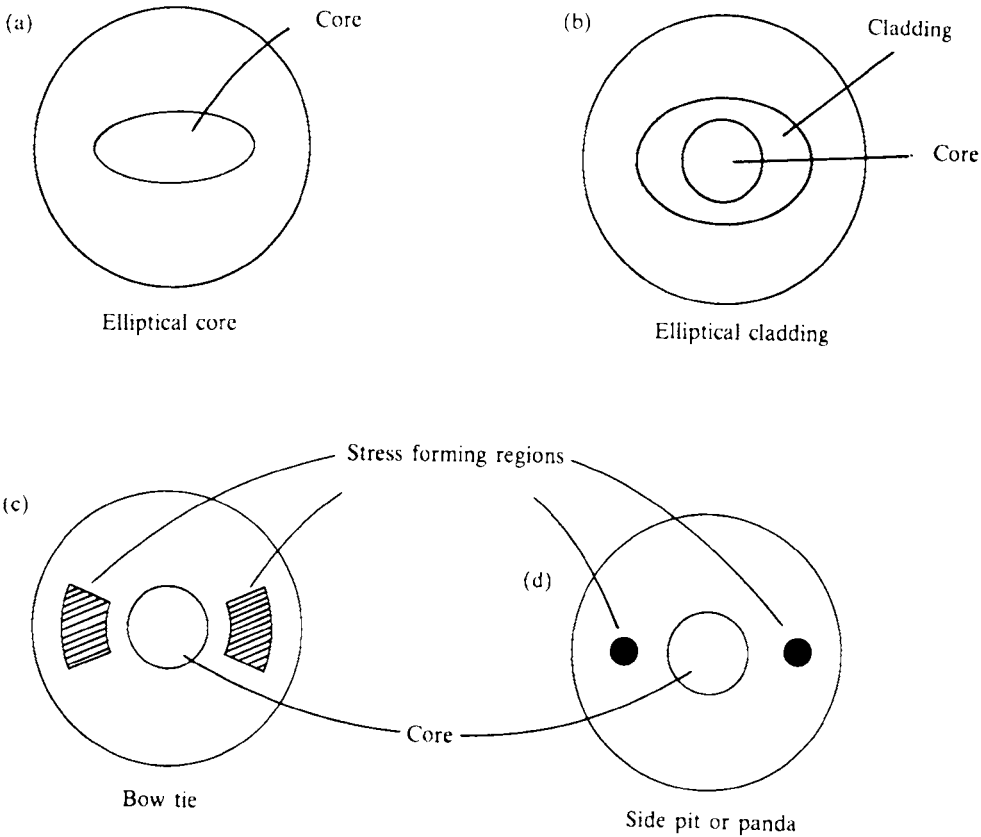


FIG. 8.52 Various types of birefringent fiber: (a) shows an elliptically cored fiber, and the others all induce stress within a circular core; (b) has an elliptical cladding whilst (c) and (d) have stress forming regions within the cladding.

are difficult to manufacture. Most Hi-Bi fibers are made by incorporating a localized stress region into the fiber. Thus an elliptical cladding (Fig. 8.52b) produces an asymmetric stress across the core which leads to birefringence. Other examples are the ‘bow tie’ fiber (Fig. 8.52c), which is fabricated using a gas phase etching process during manufacture of the preform, to produce stress regions round the core, and the ‘side pit’ or ‘panda’ fiber (Fig. 8.52d), which incorporates borosilicate rods in the cladding.

It is also possible to design a fiber such that only one of the orthogonal modes can propagate at the design wavelength. This is done by choosing the propagation constants so that one of the modes is a guided mode whilst the other can only propagate as a leaky mode, and is accordingly highly attenuated.

8.7.5.2 Bragg fiber gratings

When germania-doped fibers are exposed to ultraviolet radiation with a wavelength between 240 nm and 250 nm a change takes place in the refractive index of the fiber. The change is

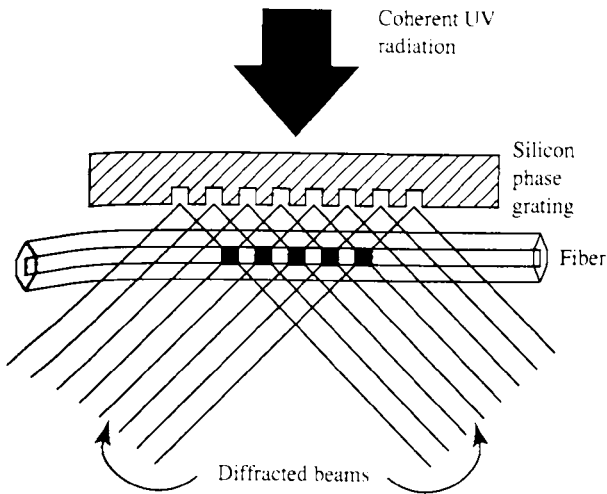


FIG. 8.53 Use of a silica phase mask to form a phase grating in a fiber. The phase mask has a wavelength D which is twice that required. The two first-order beams from the diffracted UV beam can produce a fiber grating having a periodicity of Λ , where $\Lambda = D/2$.

relatively small (typically less than 10^{-4}) but it is permanent, and can be increased by the use of a technique called *hydrogen loading*. This involves forcing molecular hydrogen into the fiber core by means of a high pressure, low temperature diffusion process. The resulting refractive index changes can be used to create a *phase grating*, that is a periodic change in the refractive index, within the core. There are a number of techniques for doing this: for example, the fiber may be illuminated from the side by a coherent ultraviolet source which is passed through a silica phase mask. This causes the beam to diffract into a number of separate beams which can be used to 'write' the required pattern onto the core of the fiber, as shown in Fig. 8.53. Typically the grating pattern is between 1 mm and 20 mm in length. Once formed the fiber grating behaves in a very similar fashion to the Bragg acousto-optic grating of section 3.8. Thus radiation of wavelength λ_0 which satisfies the equation

$$\lambda_0 = 2mn_1\Lambda \quad (8.49)$$

where m is an integer, Λ the periodicity of the grating and n_1 the refractive index of the core, will undergo strong reflection when it encounters the grating. Such a structure is very useful in a number of situations. It can provide a suitable mirror for use with fiber lasers (section 5.10.1), it can be the basis of a wavelength-selective device in optical fiber communications systems (section 9.3.8) and also as the basis for optical fiber temperature and strain sensors (section 10.1.2).

8.8

Fiber cables

The tensile strength of glass and silica fibers, when freshly made, is very high and indeed compares favourably with that of steel. However, surface damage caused by handling or even

atmospheric attack rapidly leads to a decrease in strength. This effect may be greatly reduced by adding a coating layer to the fiber immediately after manufacture. The material used must provide a good chemical and physical barrier and yet must be fairly readily removable for the purpose of jointing. The most widely used material is kynar, a vinylidene fluoride polymer commonly used as an electrical insulator. Although this primary coating helps preserve the intrinsically high tensile strength of the fiber, it cannot protect it from major mechanical damage. Consequently when fiber is used in any but a laboratory environment, it is invariably contained in some form of cable structure. Care must be taken to avoid the occurrence of microbends, since, as we saw in section 8.4.1, they can lead to significant additional fiber losses of the order of 1 or 2 dB km⁻¹.

In many cable designs, the fiber is given a further coating or contained within a tight-fitting tube to make the fibers easy to handle during cable manufacture. Care is needed to ensure that differential thermal expansion or ageing processes do not again give rise to microbends or place undue stress on the fiber. The most successful materials for this have been relatively hard ones such as nylon, polypropylene or polyurethane.

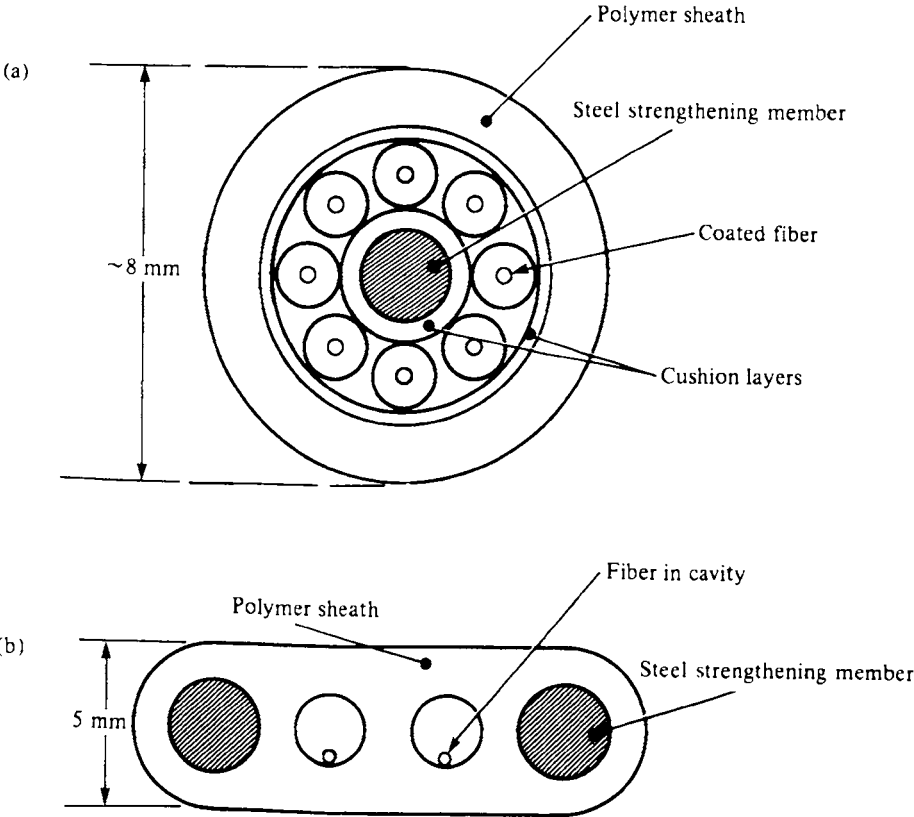


FIG. 8.54 Typical cable designs. In (a) the coated fibers are wrapped helically round a central strengthening member, while (b) shows the BICC 'PSB' design where the bare fiber lies loosely within a cavity in the cable.

When incorporated into the main cable structure, the fibers are often wrapped helically round a central strengthening member and contained within a polymer sheath (Fig. 8.54a). Designs have also been proposed where the second fiber coating is dispensed with, and the fiber lies loosely in a cavity within the cable structure, an example of this being the BICC 'PSB' design shown in Fig. 8.54(b).

Steps are usually taken to exclude water from the cable. This may be done by incorporating an axially laid aluminium foil/polyethylene laminated film immediately inside the outer sheath. Alternatively the spaces within the cable may be filled with a moisture-resistant silicone rubber or petroleum-based compounds. The exclusion of water is important in that it has been found that contact with water causes an increase in attenuation in silica fibers. When contact occurs diffusion of hydrogen into the fiber takes place. It either moves into interstitial spaces in the structure or reacts with the fiber dopants to form P-OH , Ge-OH or Si-OH . Increases in fiber absorption of 5 dB km^{-1} or more can result. Further details on fiber cable design are contained in ref. 8.27.

NOTES

1. It is perhaps unfortunate that the commonly used symbol for this quantity should be identical for voltage, but there should be no confusion in the context in which it is used here.
2. In fact, there will be two modes, TE_0 and TM_0 .
3. It will be noted that we only have considered the case where $\theta < 45^\circ$ in Fig. 8.10. It is left as an exercise for the reader to draw an appropriate diagram for the case of $\theta > 45^\circ$ and to show that the corresponding phase difference between the intersecting beams is the same as in the $\theta < 45^\circ$ case.
4. In circular waveguides the core field amplitudes are written in terms of Bessel functions; the number 2.405 represents the first zero of the $J_0(x)$ Bessel function. We may note that the equivalent condition for planar waveguides is $V = \pi/2$. Significantly this represents the first zero of the cosine function, which is just the function which describes the core field amplitudes in such guides.

PROBLEMS

- 8.1 Calculate the critical angle for a water/air interface. The refractive index of water is 1.33.
- 8.2 Verify that the phase change δ on reflection at an interface for radiation polarized with its electric vector parallel to the plane of incidence is given by eq. (8.6a).
- 8.3 Calculate the phase shifts introduced on reflection inside a planar waveguide where $n_1 = 1.48$, $n_2 = 1.46$ and the angle of incidence is 85° .
- 8.4 Light of vacuum wavelength $0.85 \mu\text{m}$ travelling in a medium of refractive index 1.48 is incident upon an interface with another medium of refractive index 1.46 at an angle

of 82° . Calculate the relative amounts of power present at distances (a) $5\text{ }\mu\text{m}$ and (b) $50\text{ }\mu\text{m}$ into the second medium.

- 8.5 Sketch the graphs (to scale) required to determine the ray angles of the allowed modes in a planar symmetric waveguide of thickness $4\text{ }\mu\text{m}$ where the core and cladding refractive indices are $n_1 = 1.50$ and $n_2 = 1$ respectively and where radiation of wavelength $1\text{ }\mu\text{m}$ is being used.
- 8.6 Estimate the attenuation of a 'non-guided' ray down a planar dielectric waveguide where $n_1 = 1.48$, $n_2 = 1.46$ and $d = 200\text{ }\mu\text{m}$. Assume that the ray has an internal angle of 0.1° less than the critical angle and that the electric field is perpendicular to the plane of incidence.
- 8.7 The output from a semiconductor laser is usually linearly polarized with the electric field vector in the plane of the junction. Explain why this should be. (Hint: consider the relative phase shifts expected on reflection (Fig. 8.2) and the resulting internal beam angles together with the consequent spread of the mode energy into the cladding.)
- 8.8 A step index fiber has a numerical aperture of 0.16, a core refractive index of 1.47 and a core diameter of $200\text{ }\mu\text{m}$. Calculate (a) the acceptance angle of the fiber, (b) the refractive index of the cladding and (c) the approximate maximum number of modes with a wavelength of $0.85\text{ }\mu\text{m}$ that the fiber can carry.

If the fiber is immersed in water (refractive index 1.33) will the acceptance angle change?

- 8.9 Estimate the intermodal dispersion expected for the fiber in Problem 8.8.
- 8.10 Calculate the maximum core diameter needed for a circular dielectric waveguide (refractive index 1.5) to support a single mode of radiation with wavelength $1\text{ }\mu\text{m}$ when (a) the waveguide has no cladding and (b) the waveguide is of the step index type with $\Delta = 0.01$.

Comment on the values obtained from the point of view of manufacturing single mode waveguides.

- 8.11 A laser operating at $1.5\text{ }\mu\text{m}$ with a linewidth of $2\text{ }\mu\text{m}$ is used in conjunction with a single mode silica-based fiber. Estimate the material dispersion expected.
- 8.12 Show that when the source linewidth is exactly centred on the wavelength ($=\lambda_m$), where $\lambda_0^2 d^2 n / d\lambda_0^2 = 0$, then the resulting material dispersion can be written

$$\Delta\tau = \frac{L}{8c} \left[\lambda_m^3 \left(\frac{d^3 n}{d\lambda^3} \right)_{\lambda=\lambda_m} \right] \left(\frac{\Delta\lambda}{\lambda_m} \right)$$

(Hint: the simplest way of proceeding is via the group index N_g (eq. 8.32); at λ_m , N_g will have a minimum value and hence about λ_m will have a $(\lambda_0 - \lambda_m)^2$ dependence on λ_0 . Thus we may write $N(\lambda_0) - N(\lambda_m) = [(\lambda_0 - \lambda_m)^2 / 2!](d^2 N / d\lambda_0^2)_{\lambda_m}$.)

- 8.13 Consider a waveguide bent into an arc of a circle of radius R . Determine the point in the cladding at which the phase velocity of a guided mode equals the phase velocity

of a plane wave in the cladding. Assume that the mode phase velocity is given by c/n_1 . Hence show that the rate of energy loss is proportional to $\exp(-R/R_0)$ where R_0 is a constant. Assume that the mode electric field intensity declines exponentially with distance into the cladding.

8.14 An optical power of 1 mW is launched into an optical fiber of length 100 m. If the power emerging from the other end is 0.3 mW, calculate the fiber attenuation (ignore reflection losses).

8.15 In a particular ZBLAN type of fiber the individual contributions to the fiber attenuation were determined to be as follows:

Rayleigh scattering: $0.9 + 0.67\lambda^{-4}$ dB km⁻¹

tail from phonon absorption peak: $2 \times 10^{-3} \exp(3/\lambda)$ dB km⁻¹

Determine the minimum attenuation achieved and the wavelength at which it occurs.

8.16 Show that the relationship between the units dB km⁻¹ and absorption coefficient α (m⁻¹) is given by:

$$\text{dB km}^{-1} \equiv 4343\alpha$$

8.17 Two identical multimode fibers (of core radius a) have their end faces in contact but their core centres offset by a distance D . By ignoring Fresnel losses and assuming that the energy within each fiber is uniformly distributed across the fiber core, show that the fractional energy transmission between the fibers is given by eq. (8.40).

If the fibers have no offset displacement, but instead are separated by a longitudinal displacement S , show that the fractional coupling loss can be written as $a^2/(a + S \tan \alpha_{\max})^2$, where it has been assumed that radiation emerges from the fiber end with a uniform angular distribution up to the fiber acceptance angle, α_{\max} .

In both cases sketch your results in the form of dB loss versus normalized misalignment (i.e. S/a and D/a) and compare with the practical results shown in Fig. 8.35 (you will have to decide on a suitable value for α_{\max}).

8.18 In an experiment to determine fiber scattering loss, a fiber is passed diagonally through a cube whose sides are lined with Si photodetectors. Light from an He-Ne laser ($\lambda_0 = 633$ nm) is coupled into one end of the fiber and 4 mW of power are found to be emerging from the other (a short length is used). If the cube side length is 30 mm and the short circuit current obtained from all the cells connected in parallel is 100 nA, estimate the scattering loss in the fiber (in dB km⁻¹). You will need to assume a value for the quantum efficiency of the detectors.

8.19 A Gaussian pulse with $\sigma = 2$ ns is launched into a fiber 10 km long. At the other end, the pulse is still Gaussian but with $\sigma = 10$ ns. Estimate the (electrical) bandwidth of the fiber.

8.20 The refractive index profile of a graded index fiber is to be examined using the near-field scanning technique. Draw a schematic diagram of the apparatus required and show explicitly how the profile parameter may be obtained from the results.

- 8.21 When scattering takes place within a fiber, some of the scattered radiation will enter guided modes and travel back down the fiber. Assuming isotropic scattering within a step index fiber of numerical aperture NA and cladding refractive index n_2 , show that the fraction of the scattered radiation so captured is given, approximately, by $(NA/2n_2)^2$.
- 8.22 Deduce all you can about the fiber whose OTDR trace is shown in Fig. 8.46.
- 8.23 It has been suggested that if a material could be found where the onset of the infrared lattice absorption bands is at a higher wavelength than in glass/silica fibers, then much lower ultimate fiber losses could be obtained. Estimate from Fig. 8.32 what the minimum absorption is likely to be at a wavelength of 4 μm , assuming Rayleigh scattering is the dominant loss mechanism.

REFERENCES

- 8.1 J. Tyndall, *R. Inst. GB Proc.*, **6**, 189, 1870–2.
- 8.2 K. C. Kao and G. A. Hockham, 'Dielectric-fibre surface waveguides for optical frequencies', *Proc. IEEE*, **113**, 1151, 1966.
- 8.3 A. C. S. van Heel, 'A new method of transporting optical images without aberrations', *Nature*, **173**, 39, 1954.
- 8.4 M. Born and E. Wolf, *Principles of Optics* (6th edn), Pergamon, Oxford, 1980, Section 1.5.2.
- 8.5 A. H. Cherin, *An Introduction to Optical Fibers*, McGraw-Hill, New York, 1983, Chapter 4.
- 8.6 J. E. Midwinter, *Optical Fibers for Transmission*, Wiley-Interscience, New York, 1979, Chapter 5.
- 8.7 H.-G. Unger, *Planar Optical Waveguides and Fibers*, Oxford University Press, Oxford, 1977, Chapter 5.
- 8.8 D. Gloge and E. A. J. Marcatili, 'Multimode theory of graded-core fibers', *Bell Syst. Tech. J.*, **523**, 1563, 1973.
- 8.9 G. Einarsson, *Principles of Lightwave Communications*, John Wiley, Chichester, 1996, Section 2.3.
- 8.10 Luc B. Jeunhomme, *Single Mode Fiber Optics* (2nd edn), Marcel Dekker, New York, 1990, pp. 17–20.
- 8.11 J. E. Midwinter, *op. cit.*, Appendices 1, 2 and 3.
- 8.12 W. A. Gardener, 'Microbending loss in optical fiber', *Bell Syst. Tech. J.*, **54**, 457, 1975.
- 8.13 E.-G. Neumann and H.-D. Rudolph, 'Radiation from bends in dielectric transmission lines', *IEEE Trans. Microwave Theor. Technol.*, **MITT-23**, 142–9, 1975.
- 8.14 R. Olshansky, 'Propagation in glass optical waveguides', *Rev. Mod. Phys.*, **51**(2), 341, 1971.
- 8.15 T. Miya, Y. Terunuma, T. Hosaka and T. Miyashita, 'Ultimate low-loss single-mode fiber at 1.55 μm ', *Electron. Lett.*, **15**, 106, 1979.

- 8.16 S. Nemeto and T. Makimoto, 'Analysis of splice loss in single-mode fibers using a Gaussian field approximation', *Opt. Quantum Electron.*, **11**, 447–57, 1979.
- 8.17 J. Senior, *Optical Fiber Communications, Principles and Practice* (2nd edn), Prentice Hall International, Hemel Hempstead, 1992, Sections 5.4 and 5.5.
- 8.18 J. E. Midwinter, *op. cit.*, Chapters 10, 11.
- 8.19 J. E. Midwinter, *op. cit.*, Section 8.3.
- 8.20 J. Gowar, *Optical Communication Systems* (2nd edn), Prentice Hall, Hemel Hempstead, 1993, Section A1.2.
- 8.21 A. H. Cherin, *op. cit.*, Section 6.8.
- 8.22 M. J. Adams, D. N. Payne and F. M. E. Sladen, 'Correction factors for the determination of optical fiber refractive index profiles by near field scanning techniques', *Electron. Lett.*, **125**, 281, 1976.
- 8.23 G. Keiser, *Optical Fiber Communications*, McGraw-Hill, London, 1983, Section 9.4.
- 8.24 S. R. Nagel, 'Fiber materials and fabrication methods', in S. E. Millar and I. P. Kaminow (eds) *Optical Fiber Telecommunications II*, Academic Press, Boston, 1988, pp. 121–215.
- 8.25 (a) J. Nishii *et al.*, 'Recent advances and trends in chalcogenide glass fiber technology: a review', *J. Non-Cryst. Solids*, **140**, 199–208, 1992.
 (b) S. Takahashi, 'Prospects for ultra-low loss using fluoride glass optical fiber: a review', *J. Non-Cryst. Solids*, **140**, 172–8, 1992.
- 8.26 J. Wilson and J. F. B. Hawkes, *Lasers: Principles and Applications*, Prentice Hall International, Hemel Hempstead, 1987, Section 5.9.
- 8.27 G. Mahlke and P. Gosling, *Fibre Optic Cables, Fundamentals, Cable Engineering, Systems Planning*, John Wiley, Chichester, 1987.
- 8.28 D. L. Lee, *Electromagnetic Principles of Integrated Optics*, John Wiley, New York, 1986, Section 8.6.

Optical communication systems

Although the benefits of an optically based communication system have long been recognized, there were two main obstacles which at first stood in the way of a successful implementation. First, until recently it was impossible to modulate and demodulate light at anything other than very low frequencies. Secondly, without some sort of guiding medium, atmospheric transmission was limited to line-of-sight communications and was subject to the vagaries of the weather. It was the development of the laser that initially aroused renewed interest in optical communications, but early systems still had to employ atmospheric transmission. However, once the initial high absorption losses were reduced, the clad dielectric waveguide became a practical proposition, and the problem of a suitable guiding medium was solved. Not all lasers are equally suitable for launching adequate power into these waveguides, and the semiconductor laser has proved the best in this respect. Interestingly, an incoherent source – the LED – is in many ways equally useful in multimode fiber systems, which brings home the point that at present the coherence properties of lasers are little used in optical communications. This is in spite of the fact that the techniques of heterodyne and homodyne mixing (section 9.1.1) do offer prospects of improved signal-to-noise ratios. At present we are still a long way from utilizing the full potential bandwidth; the highest modulation frequencies so far achieved are some 10^4 times smaller than those theoretically allowed by the carrier frequency.

In the present chapter, we review current optical communication systems and give a brief indication of the directions in which they may develop. One such area of interest in this respect is that of integrated optics. Here, the aim is to miniaturize optical components such as sources, detectors, modulators and filters and to fabricate complete optical processing systems containing these items onto a single semiconductor chip. First of all, however, we discuss the types of modulation that are most applicable to optical communications.

9.1

Modulation schemes

The term modulation describes the process of varying one of the parameters associated with a carrier wave to enable it to carry information. Variations in the amplitude, irradiance, frequency, phase and polarization can all be used for this purpose. When the carrier wave oscillates at optical frequencies, however, not all of these possibilities are equally suitable. For example, most optical detectors respond only to the irradiance of the light, and so modulation of anything other than the irradiance is not very useful unless special techniques, such as those discussed in the next section, are used.

There are many different methods for transferring information into variations of the appropriate wave parameter, and we divide these into three categories, namely *analog*, *pulse* and *digital*. With analog modulation, the primary information signal, which we take to be a time-varying electrical voltage, continuously varies the appropriate wave parameter. Thus, at any one time there is a one-to-one relationship between the original signal amplitude and the magnitude of the wave parameter. In both of the other two methods, the signal amplitude is only sampled at regular intervals and this information is then conveyed by means of a series of 'pulses'. In pulse modulation, the pulse width may be varied in proportion to the required signal; alternatively, a fixed-width pulse may be used and its time of occurrence within a fixed time slot used for the same purpose. There are also several other possibilities. In digital modulation, the information is provided by using a string of pulses whose timing and widths are both fixed but whose amplitudes are restricted to certain 'quantized' values. Figures 9.1 (a)–(c) illustrate these techniques.

Of these three categories, the most common are analog and digital, and it is to these that we now direct our attention.

9.1.1 Analog modulation

As we indicated above, not all of the wave parameters – amplitude, irradiance, frequency and phase – are equally suitable for modulation purposes at optical frequencies. The main difficulties arise in signal demodulation, since detectors respond only to the irradiance of the radiation falling on them (see eq. 7.30).

We suppose that the magnitude of the electric field of the carrier wave can be written

$$\mathcal{E}_c(t) = A_c \cos(\omega_c t + \phi_c)$$

where A_c , ω_c and ϕ_c represent the carrier amplitude, angular frequency and phase respectively, all of which may be modulated and hence be time dependent. If this signal is allowed to fall directly onto a detector (in so-called *direct* detection), then the output of a detector O_d which responds only to irradiance is given by

$$\begin{aligned} O_d &= RA_c^2 \langle \cos^2(\omega_c t + \phi_c) \rangle \\ O_d &= RA_c^2 / 2 \end{aligned} \quad (9.1)$$

Here R is the detector responsivity and the angle brackets indicate an average taken over a complete period of the function inside. Thus eq. (9.1) indicates that in direct detection only the signal irradiance (which is proportional to the square of the amplitude) is recoverable, and hence modulation of the other wave parameters is not possible. There are, however, other techniques available which do enable further information to be extracted from the received signal.

In *heterodyne* detection, the incoming signal is mixed with one from a local oscillator (Fig. 9.2) which we describe by

$$\mathcal{E}_o(t) = A_o \cos(\omega_o t + \phi_o)$$

where ω_o is chosen to be very close to ω_c , while in contrast to A_c , A_o will not, in general, be

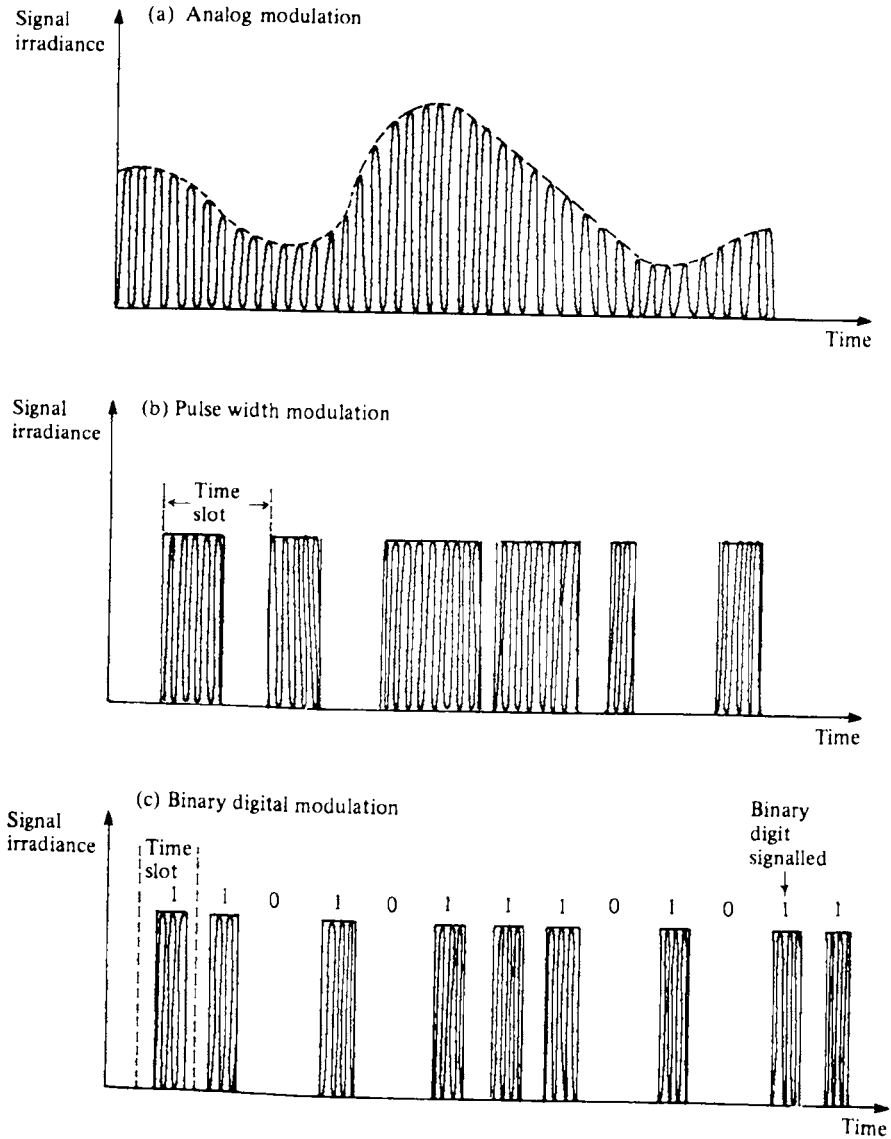


FIG. 9.1 Three possible schemes for modulating an optical light beam. In (a) the amplitude (and hence the irradiance) is continuously modulated in time. The envelope of the waveform gives the required signal. In (b) the signal is sampled at the beginning of each 'time slot' and a pulse is emitted during that time slot whose length is proportional to the signal amplitude. A digital signal is shown in (c). Here, as in (b), the signal is sampled at regular intervals and then a series of pulses is emitted which indicates (in binary notation) the signal amplitude. The presence of a pulse during a time slot indicates a 'one', while the absence of a pulse indicates a 'zero'. It should be noted that for illustrative purposes the carrier frequency has been made much smaller than would occur in practice.

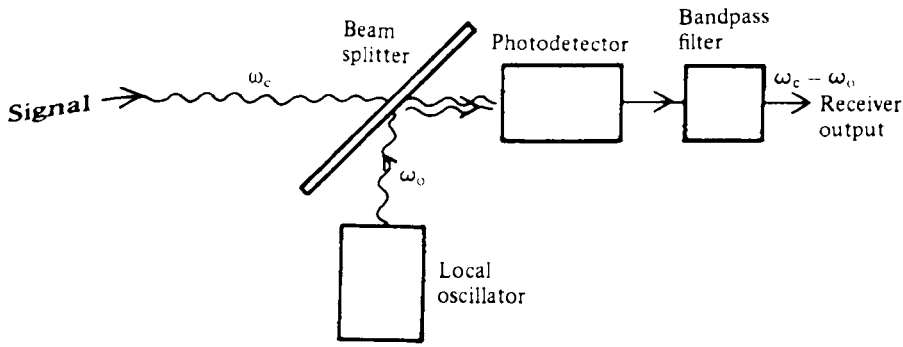


FIG. 9.2 Schematic diagram of a heterodyne detection receiver. The incoming signal (frequency ω_c) is mixed with one from a local oscillator (frequency ω_o) using a beam splitter arrangement. The combined signals fall onto a photodetector and the electrical output of the latter is passed through a bandpass filter centred on frequency $\omega_c - \omega_o$.

time dependent. The output of the detector now becomes

$$\begin{aligned}
 O_d &= R\langle [\mathcal{E}_c(t) + \mathcal{E}_o(t)]^2 \rangle \\
 &= R\langle \mathcal{E}_c^2(t) + \mathcal{E}_o^2(t) + 2\mathcal{E}_c(t)\mathcal{E}_o(t) \rangle \\
 &= R\left(\frac{A_c^2}{2} + \frac{A_o^2}{2} + 2A_cA_o\langle \cos(\omega_c t + \phi_c) \cos(\omega_o t + \phi_o) \rangle \right)
 \end{aligned} \quad (9.2)$$

Now

$$\langle \cos(\omega_c t + \phi_c) \cos(\omega_o t + \phi_o) \rangle = \frac{1}{2} \langle \cos[(\omega_c + \omega_o)t + (\phi_c + \phi_o)] + \cos[(\omega_c - \omega_o)t + (\phi_c - \phi_o)] \rangle$$

Since ω_o is close to ω_c , the term $\cos[(\omega_c - \omega_o)t + (\phi_c - \phi_o)]$ oscillates much more slowly than the other and thus it may be separated out by inserting an electrical bandpass filter centred on $(\omega_c - \omega_o)$ in the output of the detector (Fig. 9.2). The resulting output is then

$$O_d = RA_cA_o \cos[(\omega_c - \omega_o)t + (\phi_c - \phi_o)] \quad (9.3)$$

O_d is now dependent on carrier amplitude, angular frequency and phase; hence each of these parameters may be used to carry information.

In *homodyne* detection, the local oscillator is set to the *same* frequency as the carrier. In this situation, eq. (9.2) becomes

$$O_d = R\left(\frac{A_c^2}{2} + \frac{A_o^2}{2} + A_cA_o \cos(\phi_c - \phi_o) + A_cA_o \langle \cos[2\omega_c t + (\phi_c + \phi_o)] \rangle \right)$$

By inserting a low frequency (i.e. much less than ω_c) bandpass electrical filter after the detector, we may block terms other than $A_c^2/2$ and $A_cA_o \cos(\phi_c - \phi_o)$. If, in addition, $A_c \ll A_o$ then we have

$$O_d = RA_cA_o \cos(\phi_c - \phi_o) \quad (9.4)$$

Thus, in homodyne detection both amplitude and phase modulation are possible. We also

see that for both heterodyne and homodyne detection the signal amplitude is a factor A_o/A_c larger than for direct detection. When $A_o \gg A_c$, we have signal amplification, which enables higher sensitivity to be obtained. Thus A_o can be increased to the level where the receiver noise is dominated by the shot noise contribution from the local oscillator (section 9.3.3.1). In practice, improvements of some 10 to 20 dB in signal-to-noise ratios have been observed over direct detection techniques (ref. 9.1). Coherent systems are not yet, however, in common use mainly because of the additional complexity (and hence expense) required to achieve the necessary frequency, phase and polarization stability. They will be discussed further in section 9.3.9. For most of the rest of the chapter we will assume the use of direct detection.

9.1.2 Digital modulation

In digital modulation the information carrier can assume one of a number of discrete states. The simplest scheme is called *two-level binary*; in this there are only two such discrete states, which are conveniently referred to as 'zero' and 'one'. Thus in amplitude modulation the 'one' state may be represented by a pulse which is greater than some predetermined amplitude and the 'zero' state by a pulse that is less than this amplitude. The pulses are taken to occur within fixed time slots. If the pulse width is less than that of the time slot it occupies, then we refer to a *return-to-zero* (RZ) signal. Conversely, if the pulse fills the time slot, we refer to a *non-return-to-zero* (NRZ) signal. Figure 9.3 illustrates these ideas. The complete

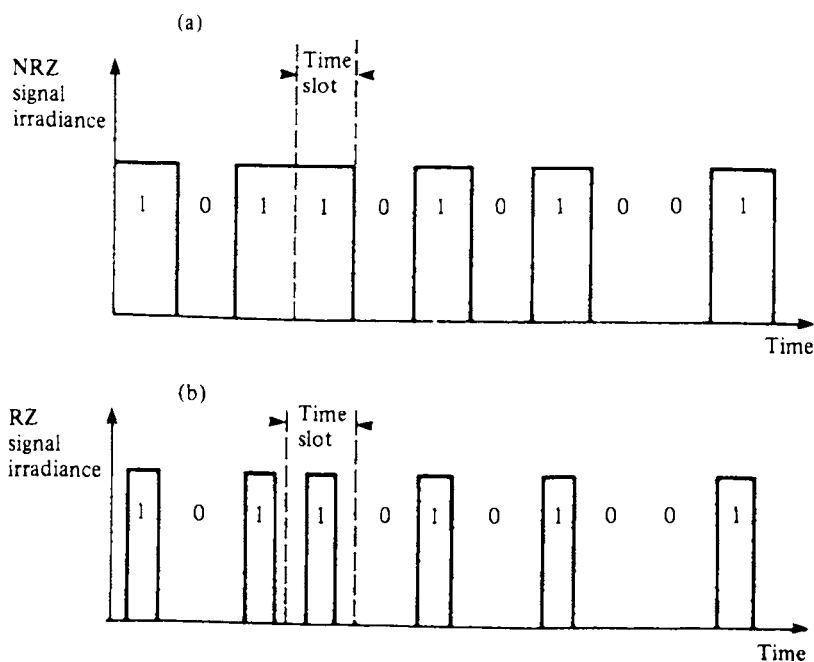


FIG. 9.3 In a non-return-to-zero (NRZ) signal (a), each pulse exactly fills the time slot. In a return-to-zero (RZ) signal (b), the pulse width is less than that of the time slot.

process whereby an analog signal is converted into a digital-pulse-coded one is shown in Fig. 9.4. The amplitude of the incoming signal is measured at discrete intervals and the result converted into a binary number. For example, in Fig. 9.4(a) the signal amplitude at $t = 2$ ms is 3 V, the number 3 in binary notation is 011 (we employ a three-digit binary number here for simplicity; in practice, eight or more binary digits or *bits* are required). During the time interval between this sampling event and the next, the information is transmitted as a series of pulses. Thus, the signal heights at the five times indicated in Fig. 9.4(a) become converted into the digital signal shown in Fig. 9.4(b).

At first sight, this scheme seems somewhat wasteful since it requires a much higher system frequency bandwidth than would be needed for the corresponding analog signal. To reproduce a given signal, the *sampling theorem* (see ref. 9.2) tells us that we must sample at a rate that is at least twice that of the highest frequency component in the signal. For example, if an 8 bit number is used for each amplitude measurement the bit rate required will be 16 times the highest signal frequency. To reproduce each bit with reasonable accuracy, the system frequency response must be at least equal to the bit rate. In telephone communication systems, for example, the highest signal frequency is usually 4 kHz; a digital system would therefore require a bit rate of some 4×16 kbps or 64 kbps and the frequency bandwidth would have to be greater than this to enable the shape of the pulses to be reproduced with reasonable accuracy.

The great advantage of a digital system over an analog one is its relative freedom from noise or distortion. Inevitably in any communication system the transmitted and received signals will not be identical. Noise may be introduced and non-linearities in component

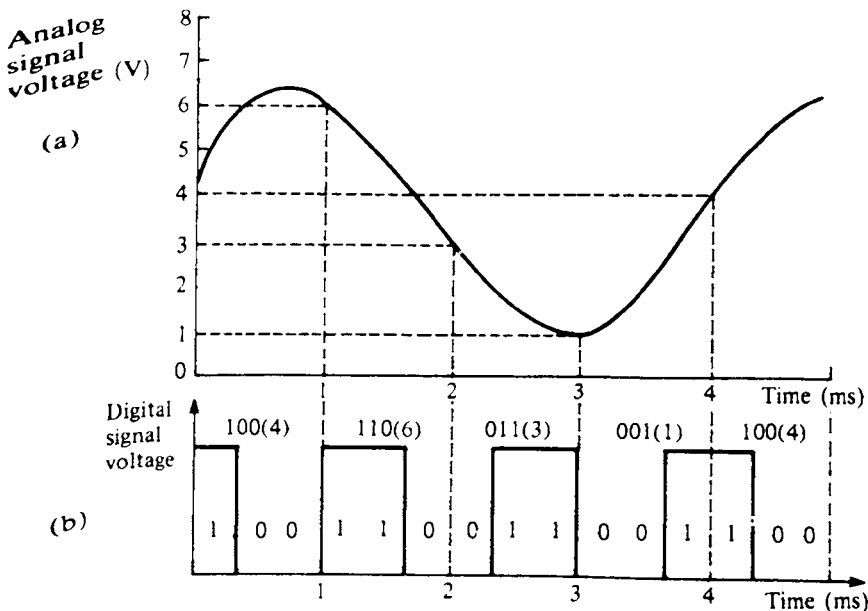


FIG. 9.4 Transformation of an analog signal into a digital one. The analog voltage (a) is sampled every millisecond and each measurement converted into a 3 bit two-level binary NRZ signal (b).

response will distort the signal. In analog systems, there is no means of rectifying this (at least when using direct detection), but a digital signal can suffer severe distortion and still be capable of accurately imparting the original information. In Fig. 9.5 we show a severely degraded signal but, provided that during each time slot we can always make the correct decision as to whether it contains a 'one' pulse or a 'zero', the original signal can be exactly reproduced. The way in which this decision is usually reached is to set up a *decision level*. If the signal exceeds this level at a particular time (the *decision time*) within each time slot, then a 'one' is recorded; if it does not do so, then a 'zero' is recorded.

This does not mean, however, that digital signals are entirely error free. Inevitably the fluctuations in signal level mean that there is always a possibility of making a mistake as to whether a 'one' or 'zero' has been received. The *bit-error rate* indicates, for some average signal level, the probability that a particular bit will be in error. If it is assumed that the signal amplitudes exhibit a Gaussian distribution, and that the decision level is set half way between the 'zero' level and the 'one' level, then it can be shown (ref. 9.3) that the relationship between the bit-error rate, BER, and the (electrical) signal-to-noise ratio in the signal, S/N , is given by

$$\text{BER} = \frac{1}{2} \operatorname{erfc} \left(\frac{(S/N)^{1/2}}{2\sqrt{2}} \right) \quad (9.5)$$

where erfc is the *complementary error function*, given by

$$\operatorname{erfc}(y) = 1 - \frac{2}{\sqrt{\pi}} \int_0^y \exp(-x^2) dx \quad (9.5a)$$

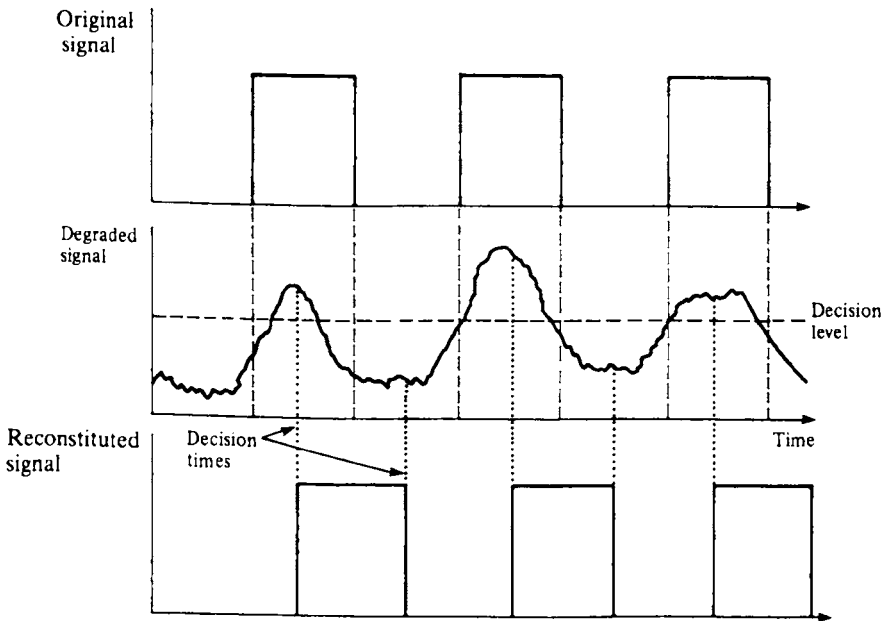


FIG. 9.5 Illustration showing how a severely degraded digital signal can still be restored to its original condition.

A graph of the BER as a function of S/N ratio is shown in Fig. 9.6. A fairly standard BER for telecommunication systems is 10^{-9} ; this is seen to require an S/N ratio of 21.6 dB (or 144).

We have assumed in this discussion that the detector 'knows' the position and spacing of the time slots and also the height of the decision level. Usually, however, the detector can only determine this information from the signal itself. A standard sequence inserted before the message proper begins helps, but both the pulse amplitude and the timing of the pulses may very well 'wander' somewhat during the message, and the detector circuitry must be able to maintain updated values of these parameters. Fortuitous groupings of pulses can make this task difficult. For example, the decision level is usually determined by taking the average amplitude of a sequence of received bits. If there are equal numbers of 'ones' and 'zeros', this procedure will be adequate. If, however, a fairly long sequence of the same bit occurs, which is not as unlikely as it seems at first, then the decision level may change to such an extent that errors become likely when a more even sequence subsequently appears. To overcome these problems, it is customary to modify the message so that sequences do contain a fairly even distribution of 'ones' and 'zeros'. This is known as *line coding*. Several schemes for this are available (see ref. 9.4), most of which involve adding extra bits to the signal and hence increasing the required system bandwidth still further. There is generally a trade-off between coding complexity (and hence the increased bandwidth required) and the accuracy of signal recovery. A simple example of such a coding scheme is that of *coded mark inversion*. In this, an input of '0' is signalled as 01 and sequential 'ones' are signalled alternatively as 00 and 11. Long runs of the same bit are then avoided and some measure of error detection is also possible since the sequence 10, when referring to an original bit, should not occur. However, the number of bits required has now been increased by a factor of two. Most line coding schemes adopted in practice increase the required bit rate by some 20%.

In our discussion of digital modulation we have so far only considered amplitude modulation (often referred to as amplitude shift keying or simply ASK). We saw in section 9.1.1

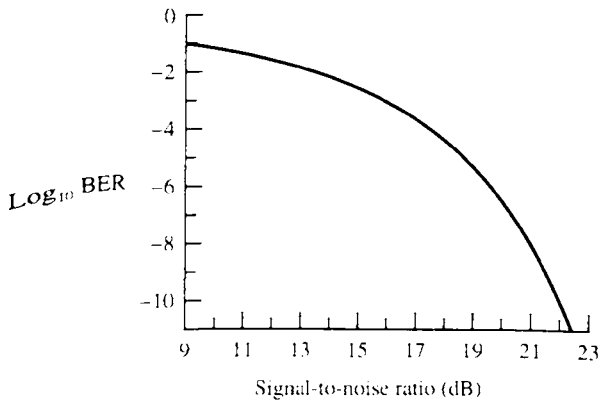


FIG. 9.6 Bit-error rate (BER) as a function of (electrical) signal-to-noise ratio as given by eq. (9.5).

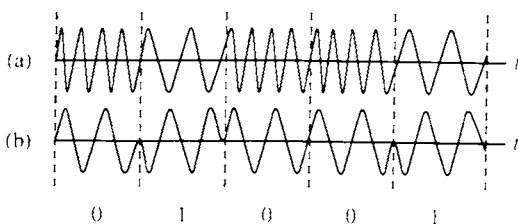


FIG. 9.7 Illustration of (a) frequency shift keying (FSK) and (b) phase shift keying (PSK). A possible interpretation of the signals in terms of 'ones' and 'zeros' is also shown.

that using coherent detection it is possible to detect changes in both frequency and phase. Thus digital modulation can also involve frequency shift keying (FSK) and phase shift keying (PSK); these are illustrated in Figs 9.7(a) and 9.7(b).

9.2

Free space communications

Although fiber optical communication has become increasingly important over the last few years, free space communication over distances of up to a few kilometres is also a practical possibility provided only fairly low bandwidths are required. Systems are easily set up and costs are considerably less than if using fibers. The low beam divergence of a well-collimated optical beam can have considerable security advantages over radio communications. A typical application would be a voice or low data rate link between nearby buildings in an urban environment.

An obvious disadvantage is that adverse atmospheric conditions, such as rain or snow, may introduce severe distortion or even render the system completely inoperable. Even when conditions are favourable, the signal will be attenuated by both absorption and scattering processes in the atmosphere. The former arises from the presence of molecular constituents such as water vapour, carbon dioxide and ozone, whose exact concentrations depend on many variables such as temperature, pressure, geographical location, altitude and weather conditions. Strong absorption occurs around wavelengths of 0.94 μm , 1.13 μm , 1.38 μm , 1.90 μm , 2.7 μm , 4.3 μm and 6.0 μm . Between these values lie the so-called *atmospheric windows*, where losses are mainly determined by scattering. Scattering is primarily due to particles that are large compared with the wavelength, such as smoke or fog, and this is known as Mie scattering (this is in contrast to Rayleigh scattering where the particle size is much smaller). The effective absorption coefficient in Mie scattering varies relatively little with wavelength. A typical atmospheric absorption spectrum is shown in Fig. 9.8.

Another problem is that of atmospheric turbulence: heating of the air in contact with the earth's surface creates convection currents in the atmosphere. Since the refractive index of air is temperature dependent, we experience refractive index variations through the atmosphere causing beam deviation and spreading. This can give rise to sudden and pronounced fading of the signal. Under ideal conditions, however, ranges of up to 150 km are possible, but to achieve reasonably large bandwidths, distances should be less than 2 km. Fluctuations

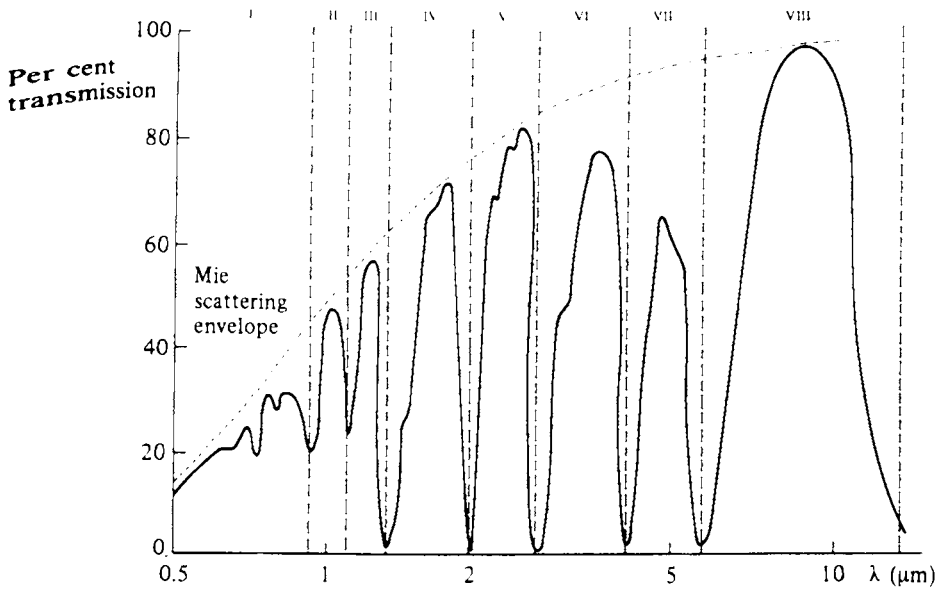


FIG. 9.8 Typical atmospheric absorption spectrum. The curve shows the percentage transmission through 1 km of atmosphere at sea level. The eight atmospheric 'windows' are indicated.

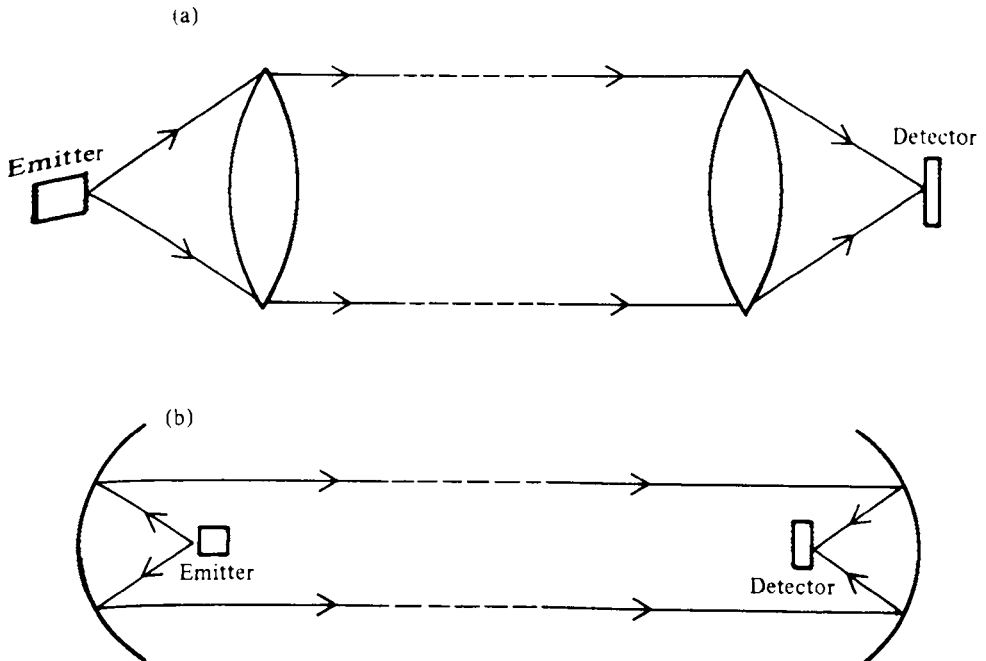


FIG. 9.9 Two simple types of optical antenna using (a) lenses and (b) concave mirrors.

in absorption and scattering tend to modulate the signal strength in the range 1–200 Hz so that amplitude modulation is not usually very satisfactory.

The simplest systems make use of LEDs and silicon photodiode detectors. These may be coupled by suitable lens or mirror systems such as those shown in Fig. 9.9. More complex systems have been proposed using sources such as Nd:YAG and CO₂ lasers where the inherently small laser beam divergence greatly simplifies beam collimation. Modulation may be achieved in a variety of ways. For example, a CO₂ laser operating in a single mode may be frequency modulated by introducing an electro-optic crystal into the resonant cavity. A signal voltage applied to the crystal changes its length and hence the optical length of the cavity. This in turn causes the single mode resonance frequency to change. Alternatively, lasers such as the Nd:YAG can be made to emit a train of pulses by the technique of mode locking (see section 6.3). A signal can then be impressed on this pulse train using an external modulator. Such systems as these, however, have been rarely used, primarily because of their complexity and high cost. A possible future use is in deep space communications, where there are no problems with transmission through an atmosphere and where the low beam divergences are useful in view of the large distances involved.

9.3

Fiber optical communication systems

The advent of optical fibers with losses of less than 1 dB km⁻¹ and with high information-carrying capacity (e.g. bandwidths >100 MHz) has meant that they have become very attractive alternatives to twisted wire or coaxial cables in many types of communication links. The main problem with the latter is that their attenuation increases as the square root of the frequency of the carrier. This arises because of the so-called skin effect, whereby, as frequency increases, the oscillating electrons are confined to an increasingly thin annulus round the outside of the metallic conductor. At a frequency of 100 MHz, for example, 9.5 mm diameter coaxial cable has an attenuation of about 20 dB km⁻¹, significantly worse than the optical losses of silica-based fiber even at a wavelength of 850 nm (see section 8.4).

A schematic diagram of an optical transmission system is shown in Fig. 9.10. The emitter is usually an LED or semiconductor laser, whilst the detector may be a p–i–n or avalanche photodiode. Also included on this diagram are what are known as *repeater* units which are present to counter the effects of fiber transmission losses and dispersion. In them the weak, and possibly broadened, optical pulses (we are assuming digital transmission) are detected

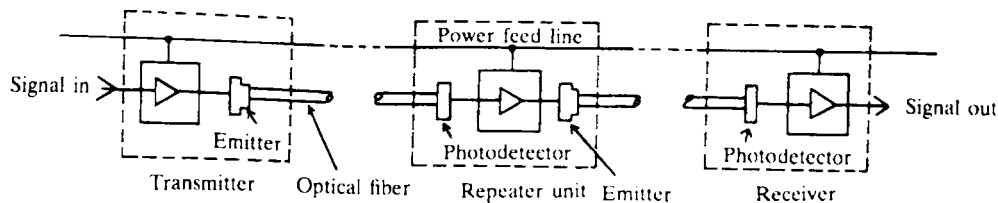


FIG. 9.10 Schematic diagram of the main components of a fiber optical communication system involving a repeater.

and converted to (weak) electrical pulses. These pulses are then reshaped, amplified and retimed, to reproduce, as far as is possible, the original noise-free pulse train. Finally a new train of optical pulses is generated for transmission down the next section of fiber. With an analog signal it is not possible to restore a distorted signal; any signal distortion or noise present is, of necessity, passed on along the next section. A separate power supply line must be provided for the repeater units and, as might be imagined, their presence adds greatly to the cost of a link. As we shall see in section 9.3.5, the need for repeaters has been almost wholly obviated by the development of the optical fiber amplifier.

The main advantages of optical fiber links include relatively low signal attenuations, high bandwidths (up to several gigahertz or more), small physical size and weight, the elimination of ground loop problems and immunity from electrical interference. This last advantage is especially useful in electrically noisy environments such as densely populated urban areas and power stations. There is, of course, no spark hazard – an additional advantage in explosive environments such as chemical plants. Fiber cables may also survive better than copper cables in certain corrosive environments, for example in sea water.

The availability of optical fibers has revolutionized telephone 'trunk' links, that is links capable of carrying a large number of simultaneous telephone conversations between telephone buildings. These may be from a few kilometres up to several hundred kilometres apart. We saw in section 9.1.2 that a single channel telephone link requires a bandwidth of 4 kHz or, for a digital signal, 64 kbps. Telephone networks require links capable of carrying many single channels simultaneously. Digital transmission is able to cope with this easily since the separate bit streams required may be interleaved. In both the United States and Europe, standard rates have been agreed upon for various 'levels' of transmission and these are shown in Table 9.1. Fiber optical links offer obvious economic advantages for medium to long haul links (10 km and upwards) with large capacities (100 Mbps and upwards).

TABLE 9.1 Digital rates used in telecommunications in Europe and the United States

Digital rates (Mbps)	Approximate number of telephone channels (1 channel = 64 kbps)	
<i>Europe:</i>		
2.048	32	
8.448	120	(4 × 30)
34.368	480	(4 × 120)
39.364	1920	(4 × 480)
565.992	7680	(2 × 7680)
1120	15 530	(2 × 7680)
2400	30 720	(2 × 15 530)
<i>United States:</i>		
1.544	T1	24
6.312	T2	96 (4 × T1)
44.736	T3	672 (7 × T2)
274.176	T4	4032 (6 × T3)

There are numerous other applications including, for example, undersea links, video transmission, computer links and, in the military sphere, missile guidance. The concept of an optical local area network (LAN) will be discussed further in section 9.3.7. Several pilot experiments have been run to examine the feasibility of providing the whole of a community's communication/information needs (such as telephone, TV, radio, etc.) using fiber optics. It is obviously impossible in a text of this size to cover these and other areas of interest adequately, and the reader is referred to ref. 9.5 for further information.

The individual components involved in a fiber optical communication link (emitters, fibers and detectors) have all been covered in general terms in previous chapters, and we now review the possibilities open to us which meet the requirements of a communication system. We consider first the choice of operating wavelength.

9.3.1 Operating wavelength

The two crucial characteristics of optical fibers that depend on wavelength are attenuation and material dispersion. A typical attenuation versus wavelength curve for a silica-based fiber was shown in Fig. 8.32. Material dispersion, that is the variation in group velocity with wavelength, was covered in section 8.3.6. There we saw (eq. 8.33) that the pulse spread $\Delta\tau$ over a fiber of length L could be written as

$$\Delta\tau \approx -\frac{L}{c} \lambda_0^2 \frac{d^2n}{d\lambda_0^2} \frac{\Delta\lambda_0}{\lambda_0}$$

The pulse spread per unit length per unit wavelength interval, $\Delta\tau/L\Delta\lambda_0$, for silica is shown in Fig. 9.11. The first-generation optical fiber transmission systems used emitters based on GaAs or GaAlAs which operate in the wavelength range 0.82 μm to 0.9 μm . This is by no means an ideal wavelength region: at 0.85 μm silica fibers have minimum attenuations of about 2 dB km^{-1} whilst their material dispersion is some 80 ps nm^{-1} . Thus, light from a typical LED with a linewidth of 50 nm would suffer a dispersion of 4 ns km^{-1} m^{-1} . Over a 10 km fiber length, the system bandwidth would then be restricted to some 25 MHz by material dispersion alone. Lasers, of course, have a considerable advantage here: by virtue of their narrow linewidths, they enable much higher bandwidths to be achieved (see Example 8.10).

Fiber attenuation may be considerably reduced by working at longer wavelengths, the minimum (about 0.15 dB km^{-1}) occurring at a wavelength of about 1.55 μm . Material dispersion also decreases at longer wavelengths, becoming very small indeed at around 1.3 μm , where values of a few picoseconds per nanometre per kilometre can be achieved. All long distance, high capacity telecommunication links now operate at wavelengths of 1.3 μm or 1.55 μm , and suitable p-i-n and APD detectors for these wavelengths based on InGaAs are now readily available.

Plastic fibers have extremely high attenuations outside the visible region (Fig. 8.41) and are usually used in short distance links in conjunction with red-emitting LEDs.

9.3.2 Emitter design

The main requirements for an emitter in an optical communication system are that it must

**Material
dispersion**
(ps nm⁻¹ km⁻¹)

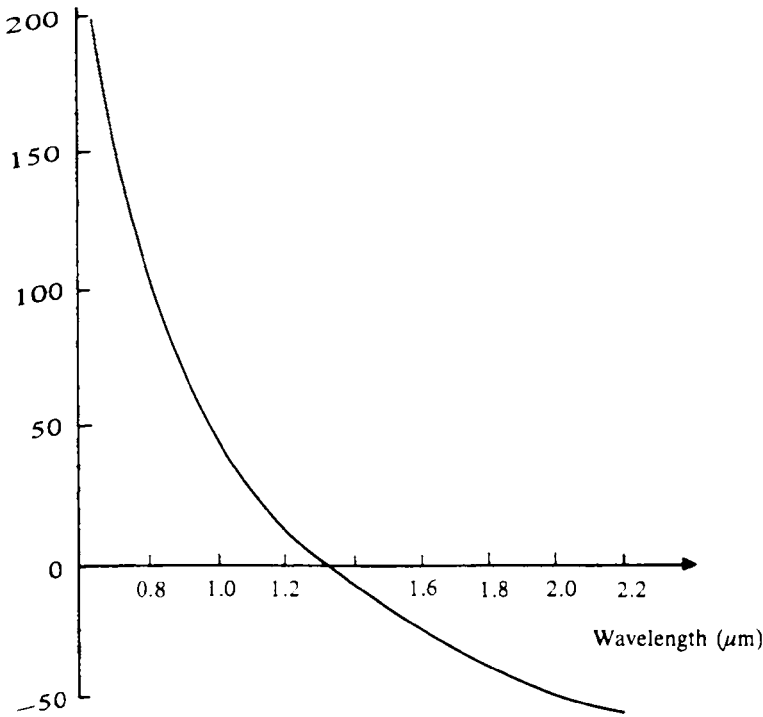


FIG. 9.11 Material dispersion for pure silica as derived from the data shown in Fig. 8.24, and expressed in ps nm⁻¹ km⁻¹.

be able to couple a useful amount of power into the fiber and also be capable of operating with the required frequency modulation bandwidth. The relatively small diameters of optical fiber cores (from several hundred micrometres down to a few micrometres) imply correspondingly small source sizes for efficient coupling of radiation into the fiber. Consequently, the most commonly used sources are LEDs and semiconductor lasers which were discussed in Chapters 4 and 5 respectively.

Compared with semiconductor lasers, LED sources are easy to drive, have long lifetimes and are inexpensive. Their principal disadvantage, apart from their greater linewidth, is that they are much less efficient at launching power into fibers than are lasers. This is mainly because of their larger emitting area, greater beam divergence and incoherent light output. We now calculate the maximum power that can be coupled into a multimode step index fiber from a source in contact with it. We assume that the source area A_s is less than or equal to the fiber core area A_c . Figure 9.12 illustrates the basic geometry. For each point on the emitting surface, only light emitted up to an angle α_{\max} , where $\alpha_{\max} = \sin^{-1}(NA/n_0)$, see eq. (8.21), with the normal to the surface will propagate down the fiber. If the source brightness as a

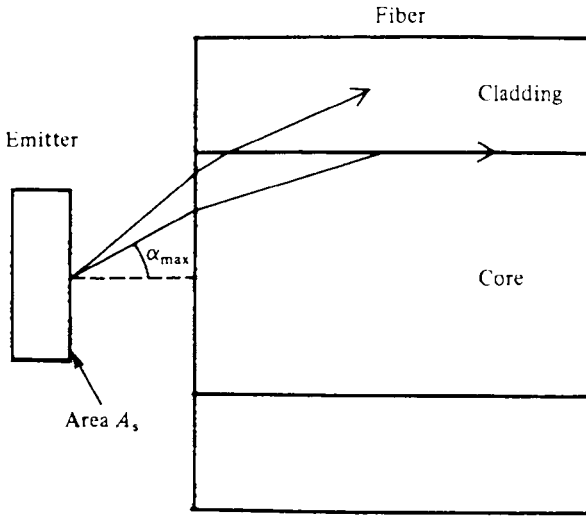


FIG. 9.12 Illustration of the coupling of light from a flat emitting surface into a fiber optical waveguide. Rays making an angle less than α_{\max} with the normal to the fiber end will become trapped in the fiber core. The gap between the emitter and the fiber has been deliberately enlarged to show the angles involved.

function of angle is written $B(\alpha)$, the total energy Φ_F coupled into the fiber is given by

$$\Phi_F = A_s \int_{\Omega} B(\alpha) d\Omega \quad (9.6)$$

Here $d\Omega$ is the solid angle subtended by rays between the angles α and $\alpha + d\alpha$, and the integral is carried out for all rays that remain trapped in the fiber. The relation between $d\Omega$ and $d\alpha$ can readily be derived with the aid of Fig. 9.13 and is

$$d\Omega = 2\pi \sin \alpha d\alpha$$

Hence, inserting this value for $d\Omega$ into eq. (9.6) yields

$$\Phi_F = 2\pi A_s \int_0^{\alpha_{\max}} B(\alpha) \sin \alpha d\alpha$$

LED sources are usually approximately Lambertian so that we may put $B(\alpha) = B(0) \cos \alpha$ (see Problem 4.6). Equation (9.6) then becomes

$$\begin{aligned} \Phi_F &= 2\pi A_s B(0) \int_0^{\alpha_{\max}} \cos \alpha \sin \alpha d\alpha \\ &= \pi A_s B(0) \sin^2 \alpha_{\max} \end{aligned}$$

or

$$\Phi_F = \pi A_s B(0) \frac{(NA)^2}{n_0^2}$$

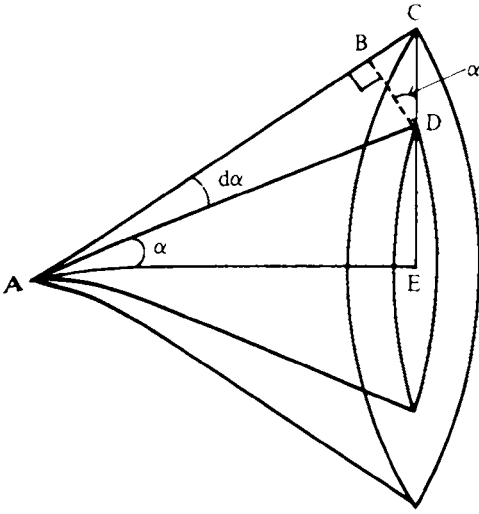


FIG. 9.13 Relation between solid angle $d\Omega$ and $d\alpha$ may be derived by noting that $d\Omega = (BD \cdot 2\pi \cdot DE)/AB^2$. Since $BD = AB \, d\alpha$ and $DE/AB \approx \sin \alpha$, we have $d\Omega = 2\pi \sin \alpha \, d\alpha$.

We know from eq. (4.21a) that the total output of the LED is $\pi A_s B(0)$ and hence the coupling efficiency η_c may be written

$$\eta_c = \frac{\text{energy coupled into fiber}}{\text{energy emitted by LED}} = \frac{\pi A_s B(0) (NA)^2}{\pi A_s B(0) n_0^2}$$

that is

$$\eta_c = \frac{(NA)^2}{n_0^2} \quad (9.7)$$

A typical value for the numerical aperture of a multimode fiber would be 0.3. Assuming an air interface (i.e. $n_0 = 1$), eq. (9.7) then shows that only some 9% at most of the total radiation emitted by the surface in contact with the core area will enter the fiber. (Fresnel losses have been ignored in this derivation since they amount to only a few per cent of the energy coupled into the fiber.) If A_s is larger than A_c , the coupling efficiency will be further reduced by a factor A_c/A_s .

For a graded index fiber the calculation is more complicated. We note, however (section 8.3.3), that a parabolically graded fiber (i.e. $\alpha = 2$) with the same diameter as a step index fiber is only able to support about half the number of modes, and hence the coupled energy is expected to be about a factor of two smaller.

To maximize the coupling efficiency into multimode fibers, we therefore require a fiber with as large an NA as possible and a source with an area no larger than the fiber core area. Increasing the NA, however, implies that the factor $(n_1 - n_2)$ increases, thus incurring the penalty of a smaller signal bandwidth because of increasing mode dispersion (see eq. 8.24).

In addition reducing the emitting area of an LED whilst maintaining total output tends to reduce the operating lifetime.

The two main types of LED most often used in fiber optical systems are the surface-etched well emitter ('Burrus' type) and the edge emitter, both of which are illustrated in Fig. 9.14. In the former, a well is etched into the top of a planar LED structure to enable the fiber end to be as close as possible to the light-emitting region (i.e. to the p-n junction). If the emitting area is less than that of the fiber core, then some form of optical coupling between the source and the fiber may be advantageous. For example, a spherical lens on top of the emitting surface will magnify the effective surface area, but demagnify the solid angular distribution of the radiation, so that the resulting radiation pattern may be a better match to the fiber acceptance pattern (Fig. 9.15). Other types of lenses that can be used to enhance source-to-fiber coupling efficiency are the sphere and GRIN types discussed in Chapter 1 (see section 1.3.3).

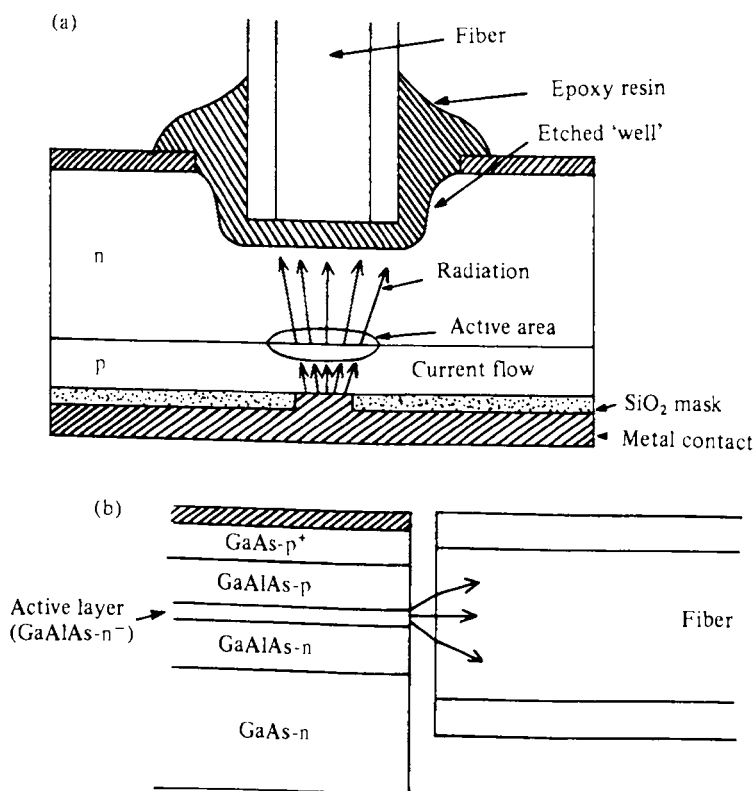


FIG. 9.14 Two types of emitter designed for more efficient coupling into optical fibers: (a) the etched well or 'Burrus' type and (b) an edge emitter. In (a), the active light-emitting area is restricted to a small region just below the end of the fiber by the use of an SiO_2 mask. The fiber is held in position by the use of a transparent epoxy resin which also helps to reduce Fresnel losses. In the edge emitter, the radiation is confined to a narrow light-guiding layer with a structure very similar to that of the double heterostructure laser.

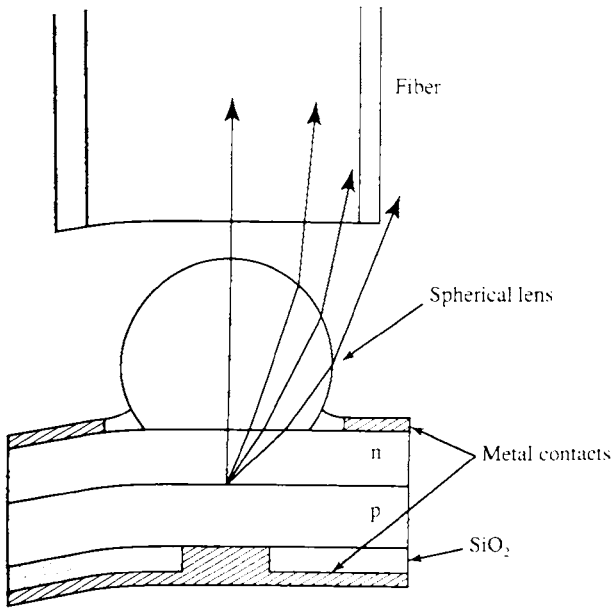


FIG. 9.15 One way devised to increase the coupling efficiency of emitters to fibers is to incorporate some type of lens system; here, the effect of a spherical lens is illustrated.

The construction of the edge-emitting LED is very similar to that of the semiconductor laser (see section 5.10.2), the radiation being confined to a narrow channel. On emerging, the beam divergences tend to be narrower than those of the surface types with typical half-power divergences of 50° and 30° perpendicular and parallel to the junction plane respectively (a Lambertian source has a half-power beam divergence of 60°). To couple this distribution into a fiber efficiently, an anamorphic optical system is sometimes used, such as that shown in Fig. 9.16.

Total optical powers from edge emitters are typically several times smaller than from surface emitters, but the narrower beam divergence can give rise to more coupled power. Edge emitters are usually preferred for use with small NA fibers (i.e. $NA < 0.4$), whereas surface emitters, because of their greater total power output, are better for large NA fibers.

As far as output characteristics are concerned, LEDs have an almost linear relationship between drive current and light output. This makes the LED more suitable for amplitude modulation, since for the latter the drive current must be switched from high to low values to obtain a wide ratio between 'on' and 'off' outputs (say from 300 mA to below 50 mA). The factors that ultimately affect the modulation bandwidths attainable are discussed in section 4.6.5. At low current levels, the limiting factor is usually the junction capacitance, whilst at high current levels it is the lifetime of carriers injected into the recombination region. It is shown in Appendix 3 that when the response is lifetime limited, we may write

$$R(f) = \frac{R(0)}{(1 + 4\pi^2 f^2 \tau^2)^{1/2}} \quad (9.8)$$

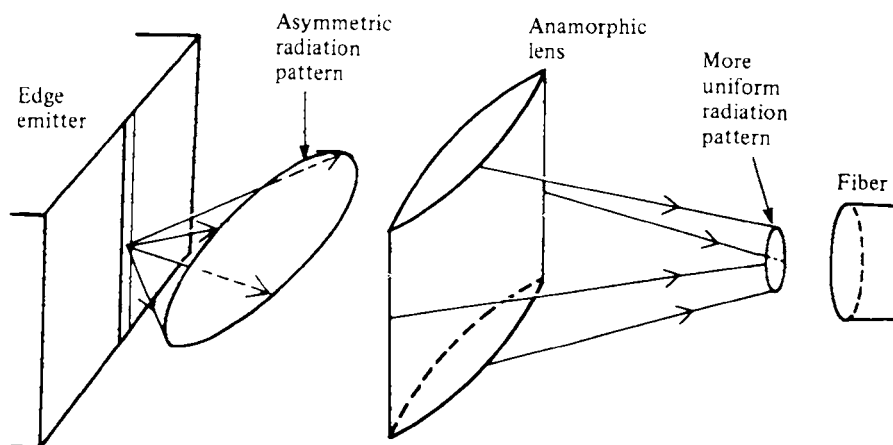


FIG. 9.16 Strongly asymmetrical radiation patterns from edge emitters can more efficiently be coupled into a fiber using an anamorphic lens (i.e. one with differing curvatures in two orthogonal directions).

where $R(f)$ is the relative response at frequency f and τ is the minority carrier lifetime. In GaAs a typical value of τ would be 10^{-9} s, giving a half-power bandwidth of some 300 MHz.

We turn now to a consideration of laser sources. In comparison with LEDs, they exhibit much higher output power, narrower beam divergence and a smaller emitting area. All of these factors enable lasers to couple significantly more power into fibers than do LEDs (see Problem 9.5). However, an even more significant factor is that lasers emit coherent radiation and it is this that enables them to couple significant power into single mode fibers. The derivation of eq. (9.7) used simple ray theory which we know is not valid in single mode fibers. To launch light into a particular mode in a fiber with 100% efficiency it is necessary in fact to reproduce on the face of the fiber the exact field amplitude distribution corresponding to that particular mode. In the event that this is not possible, then the actual power coupling efficiency into the fiber, η_c , is given by evaluating the so-called *overlap integral*, so that the coupling efficiency becomes

$$\eta_c = \frac{[\int_{-\infty}^{\infty} \mathcal{E}_m(x, y) \mathcal{E}_s(x, y) dx dy]^2}{\int_{-\infty}^{\infty} \mathcal{E}_m^*(x, y) \mathcal{E}_m(x, y) dx dy \int_{-\infty}^{\infty} \mathcal{E}_s^*(x, y) \mathcal{E}_s(x, y) dx dy} \quad (9.9)$$

where $\mathcal{E}_m(x, y)$ is the required mode field inside the fiber, and $\mathcal{E}_s(x, y)$ the actual field distribution over the end of the fiber.

As we saw in section 8.3.4 a reasonable approximation to the mode field distribution for a single mode fiber is given by a Gaussian function (eq. 8.29) and we may write the field as $\mathcal{E}_0(r) = \mathcal{E}_0(0) \exp[-(r/\omega_0)^2]$. Similarly the beam from a single mode He-Ne laser, for example, also has a Gaussian field profile. If such a beam is focused onto the end of the fiber the field profile will remain Gaussian, and we may write for the focused beam

$$\mathcal{E}_B(r) = \mathcal{E}_B(0) \exp[-(r/\omega_B)^2]$$

Equation (9.9) may now be used to evaluate the coupling efficiency of the beam into the

fiber. The integrals may be carried out using radial coordinates (see Problem 9.6), the result being

$$\eta_c = \frac{4\omega_0^2\omega_B^2}{(\omega_0^2 + \omega_B^2)^2} \quad (9.10)$$

For example, if the focused spot on the end of a single mode fiber has a mode field diameter which differs by a factor of two from that of the mode field in the fiber then from eq. (9.10) the coupling efficiency will be $(4/5)^2$ or 0.64. It is reasonably easy to obtain coupling efficiencies between 0.5 and 0.6 using simple butt jointing techniques in conjunction with single mode fibers and semiconductor laser diodes.

Lasers, of course, have very different output characteristics from LEDs. From Fig. 5.25 we see that very little radiation is emitted until the current reaches a threshold value, after which the output rises very sharply. Such a characteristic is well suited to generating digital signals, since the drive current needs only a small swing to provide a high 'on' to 'off' contrast. Analog signals can also be transmitted, since the characteristic above threshold is reasonably linear. (Sometimes, however, 'kinks' can develop as discussed in section 5.10.2.3.) One problem which severely affects both types of signal is that the threshold current is temperature sensitive and can show long term drift with age. Good temperature stabilization is therefore required together with some form of feedback mechanism to counter any change in the characteristic. The most common technique with digital signals is to maintain a constant power output. For this purpose, a photodetector may be mounted next to the rear facet of the laser. The drive current may then be adjusted until the photodetector (which need not have a very fast response) indicates some predetermined average power level.

An important requirement for high bandwidth systems is to have as small a source linewidth as possible so that material dispersion is reduced (eq. 8.34 and Example 8.10). A semiconductor laser with a Fabry-Perot-type cavity usually has a (multimode) linewidth of about 3 nm. This linewidth can be reduced by ensuring that only one longitudinal mode is present, for example by using the distributed feedback structure (section 6.2). Linewidths can then approach 10 MHz or so (corresponding to 10^{-4} nm). Further reductions (down to a few tens of kilohertz) can be obtained by using some form of external cavity (section 6.2).

Another factor that needs to be taken into account is the fact that the output pulse from a laser is 'chirped', that is the emission frequency changes during the pulse. This arises from a number of factors: for example, within the semiconductor laser material the operating wavelength depends on the product of cavity length multiplied by the refractive index. Both these quantities are affected by the changes in temperature that will take place during a current pulse. In addition the refractive index decreases as the carrier concentration increases. Thus the effective spread of wavelengths emitted by the laser is increased. This in turn increases the material dispersion and reduces the maximum bit rate possible. One solution to this is to operate the laser CW and modulate the radiation externally (possible devices will be described in section 9.4.2).

In conclusion it is evident that lasers have considerable advantages over LEDs from the point of view of coupling the maximum amount of power into fibers (especially single mode fibers) and giving minimal material dispersion. The disadvantages are that they are more expensive and require somewhat elaborate temperature control and output stabilization.

9.3.3 Detector design

Most detectors have large sensitive areas compared with fiber core areas and, Fresnel losses apart, can easily collect most of the radiation being carried by the fiber. On the other hand, too big a detector area is often a positive disadvantage since this implies an unnecessarily high dark current noise and a high detector capacitance. Thermal detectors are, in general, much too slow and insensitive and the almost universal choice has been for some kind of junction device. Of these, the main choice has been between p-i-n and avalanche photodiodes. The latter provide a substantial amount of gain which is useful in increasing the system sensitivity but only when this is limited by Johnson noise in the load resistor or noise in the amplifier following the detector. To illustrate this point we now carry out a simplified noise analysis for both a p-i-n and an avalanche (APD) detector.

9.3.3.1 Noise analysis for a p-i-n detector

The equivalent circuit of the detector and amplifier as far as noise analysis is concerned is shown in Fig. 9.17. If optical power P at a wavelength λ_0 falls on the detector then, if the detector quantum efficiency is η , the resulting signal current i_λ may be written (see eq. 7.25)

$$i_\lambda = \frac{\eta P e \lambda_0}{hc} \quad (9.11)$$

Also flowing through the load resistor is the diode reverse bias saturation, or dark, current i_D . The total current $i_\lambda + i_D$ will give rise to shot noise in the current with an r.m.s. magnitude given by eq (7.12), that is

$$\Delta(i_\lambda + i_D)_{\text{shot}} = [2(i_\lambda + i_D)e\Delta f]^{1/2} \quad (9.12)$$

Johnson noise is also present in the load resistor R_L ; from eq. (7.14) the r.m.s. Johnson noise voltage is

$$\Delta V_j = (4kTR_L\Delta f)^{1/2}$$

The equivalent r.m.s. noise current Δi_j is then

$$\Delta i_j = \frac{(4kTR_L\Delta f)^{1/2}}{R_L}$$

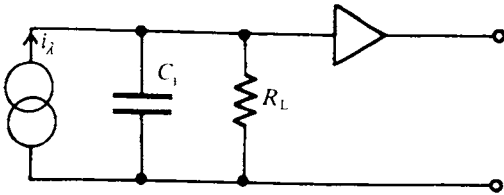


FIG. 9.17 Assumed equivalent circuit for noise analysis in a p-i-n diode. A current source i_λ is shunted by the diode junction capacitance C_i and the load resistor R_L .

Inevitably, the signal levels are usually such that amplification is required and any noise generated in the amplifier may also be important. It is useful to deal with this by including it in the Johnson noise by introducing a 'noise factor' F_n (where $F_n \geq 1$) so that

$$\Delta i_J = \frac{(4kTF_n R_L \Delta f)^{1/2}}{R_L} \quad (9.13)$$

Because the two noise terms are independent, the total noise power present within the resistor is proportional to the sum of the squares of the shot noise current and Johnson noise current. The signal power is just proportional to the square of i_λ , so that we may write the signal-to-noise power ratio as

$$S/N = \frac{i_\lambda^2}{\Delta(i_\lambda + i_D)_{\text{shot}}^2 + \Delta i_J^2}$$

or

$$S/N = \frac{i_\lambda^2}{2(i_\lambda + i_D)e\Delta f + 4kTF_n \Delta f / R_L} \quad (9.14)$$

It is evident that the importance of the Johnson noise term may be reduced, and the S/N value increased, by increasing the value of the load resistor. However, we must remember that the detector capacitance is in parallel with the load resistor and, unless some form of frequency equalization is introduced, the bandwidth will be limited to $(2\pi R_L C_j)^{-1}$ (see eq. 7.35). Thus with a bandwidth of Δf , the maximum value of R_L that may be used is given by

$$(R_L)_{\text{max}} = (2\pi C_j \Delta f)^{-1} \quad (9.15)$$

Thus the maximum S/N value is given by

$$(S/N)_{\text{max}} = \frac{i_\lambda^2}{2(i_\lambda + i_D)e\Delta f + 8\pi kTF_n \Delta f^2 C_j} \quad (9.16)$$

EXAMPLE 9.1 S/N ratio in a p-i-n photodiode

We consider a p-i-n photodiode which has a quantum efficiency of 0.6 at a wavelength of $1.3 \mu\text{m}$ and a reverse bias leakage current of 3 nA . It is used in a simple photoconductive bias circuit with a load resistor of 50Ω , the system bandwidth is 500 MHz and $10 \mu\text{W}$ of optical power (at $1.3 \mu\text{m}$) falls onto the detector. The photogenerated current (i_λ) is given by

$$i_\lambda = \frac{\eta P e \lambda_0}{hc} = \frac{0.6 \times 10 \times 10^{-6} \times 1.6 \times 10^{-19} \times 1.3 \times 10^{-6}}{6.6 \times 10^{-34} \times 3 \times 10^8} = 6.29 \mu\text{A}$$

The total shot noise current is given by eq. (9.11) as

$$\Delta(i_\lambda + i_D)_{\text{shot}} = [2(6.29 \times 10^{-6} + 3 \times 10^{-8}) \times 1.6 \times 10^{-19} \times 500 \times 10^6]^{1/2} = 31.8 \text{ nA}$$

whilst the Johnson noise current is given by eq. (9.13) as

$$\Delta i_1 = \frac{(4 \times 1.38 \times 10^{-23} \times 300 \times 50 \times 500 \times 10^6)^{1/2}}{50} = 407 \text{ nA}$$

We see that in this example the Johnson noise term is significantly larger than the shot noise term. The resulting S/N value (assuming an excess noise factor of unity) is then given by

$$S/N = \frac{(6.29 \times 10^{-6})^2}{(31.8 \times 10^{-9})^2 + (407 \times 10^{-9})^2} = 237 \text{ or } 23.7 \text{ dB}$$

However, depending on the value of the capacitance of the device this might not be the highest possible S/N value. Suppose the capacitance is 1 pF; the optimum load resistance is then given by eq. (9.15) as

$$(R_L)_{\max} = (2\pi \times 1 \times 10^{-12} \times 500 \times 10^6)^{-1} = 318 \Omega$$

With this value of R_L the Johnson noise current is reduced to 161 nA, and the signal-to-noise ratio increases to 1463 (or 31.7 dB).

Example 9.2 gives typical calculations associated with eq. (9.16). As shown in this example it is often true that the Johnson noise term is significantly larger than the shot noise. If we neglect the shot noise term in eq. (9.16) and also use eq. (9.11) to substitute for i_λ then we obtain

$$(S/N)_{\text{Johnson}} = \frac{e^2}{8\pi k h c} \frac{\eta^2 P^2 \lambda_0^2}{TF_n \Delta f^2 C_j} \quad (9.17)$$

If, however, we wish to consider the absolute limitations inherent in direct detection then we may assume that we use a sufficiently large value of R_L that the shot term dominates the denominator in eq. (9.14) (and that we may somehow compensate for the bandwidth problems). Thus in the shot-noise-limited situation and assuming that the dark current may also be neglected

$$(S/N)_{\text{shot}} = \frac{\eta P \lambda_0}{2 h c \Delta f} \quad (9.18)$$

9.3.3.2 Noise analysis for an avalanche photodiode

If the gain of the APD is M , the signal current will be $M i_\lambda$. We might then also expect that the shot noise terms would be given by replacing $(i_\lambda + i_D)$ in eq. (9.12) by $M(i_\lambda + i_D)$. However, we must remember that the avalanche process itself introduces excess noise, which is allowed for by introducing an excess noise factor $F(M)$ (see eq. 7.38). Consequently, we must write

$$\Delta(i_\lambda + i_D)_{\text{shot}} = M[2(i_\lambda + i_D)e\Delta f F(M)]^{1/2}$$

The S/N ratio then becomes

$$S/N = \frac{M^2 i_s^2}{M^2 F(M) 2e \Delta f (i_s + i_D) + 8\pi k T F_n \Delta f^2 C_j} \quad (9.19)$$

If the factor $F(M)$ were unity, the S/N ratio would increase with increasing M and would attain a maximum asymptotic value equal to that given by eq. (9.18) for shot noise alone. However, because the factor $F(M)$ increases with increasing M , the optimum S/N value is somewhat smaller than this. Figure 9.18 shows how the S/N value varies with APD gain for two values of the parameter r which is involved in the function $F(M)$ (r is the ratio of electron-to-hole ionization probabilities, see eq. 7.38a). Problem 9.7 considers the determination of the optimum APD gain in more detail.

It should be noted that the APD only improves the S/N ratio if the dominant noise term is other than shot noise. When only shot noise is present eq. (9.19) reduces to $S/N = i_s / (F(M) 2e \Delta f)$ and the S/N ratio is then a factor $F(M)$ smaller than in a p-i-n photodiode.

9.3.3.3 Fundamental limitations on signal size

Although in a practical situation Johnson noise may often be the limiting factor, it is of

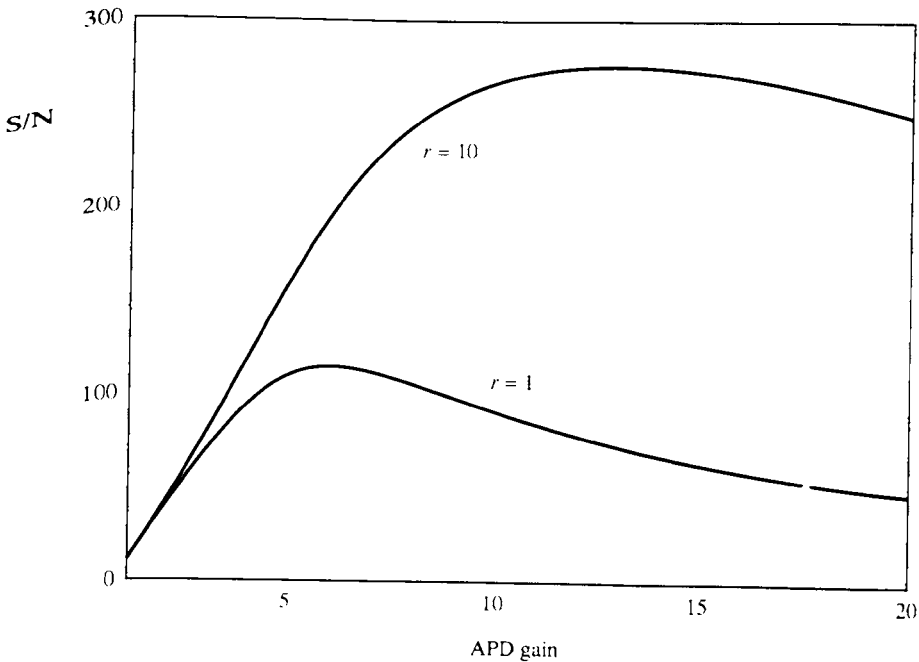


FIG. 9.18 The signal-to-noise ratio from an avalanche photodiode as a function of gain as given by eq. (9.18). Two curves are shown each corresponding to differing values of the electron-hole ionization probability (r). In the absence of any gain the detector would exhibit a Johnson noise S/N ratio of 10 and a shot noise S/N ratio of 1000. We see that as r increases the optimum S/N ratio gets closer to that arising from shot noise alone.

interest to see what the theoretical fundamental limitations on the pulse size actually are. In a digitally coded optical signal the smallest conceivable size corresponds to the arrival of a single photon. On this scale, however, the statistical nature of the photon emission and absorption processes become evident. For example, suppose the average power transmitted corresponds to 20 photons per pulse; then it is possible for pulses to arrive at the detector containing perhaps 18 or 24 or indeed any other number of photons *including zero*. It can be shown (ref. 9.6) that the process is governed by *Poisson statistics* and that the probability $p(n, n_m)$ of detecting n photons per unit time interval when the mean arrival rate is n_m is given by

$$p(n, n_m) = \frac{(n_m)^n}{n!} \exp(-n_m) \quad (9.20)$$

For example, with a mean arrival rate of 20 photons per pulse, the probability that a given pulse contains no photons at all is given by

$$p(0, 20) = \frac{(20)^0}{0!} \exp(-20) \approx 2 \times 10^{-9}$$

Consequently if we have a signal consisting of equal numbers of 'ones' and 'zeros', where the 'ones' correspond to pulses containing on average 20 photons and the 'zeros' to the arrival of no photons, then there is a probability of 2×10^{-9} that when a 'one' should be present, in fact no photons at all arrive and the pulse is consequently mistaken for a 'zero'. No error can be made in any of the 'zeros' since there can be no fluctuations in zero photons. Hence, including both 'ones' and 'zeros', a signal containing an average of 10 photons per bit will give rise to a BER of 10^{-9} . With fewer than an average of 10 photons per bit the BER will be worse (i.e. larger than 10^{-9}); conversely, with more than 10 photons per bit the BER will be better (i.e. smaller than 10^{-9}). A BER of 10^{-9} has, in fact, become a fairly standard requirement for optical communication systems. It should be noted that in the presence of even a small amount of noise the average number of photons required per bit increases sharply (ref. 9.7); for example, with a noise signal equivalent to only one photon per bit the number of signal photons per bit required to achieve a BER of 10^{-9} increases from 10 to about 22.

Remarkably perhaps, practical systems can approach to within an order of magnitude or so of this fundamental limit of 10 photons per bit. That is, they can achieve a BER of 10^{-9} whilst operating with an average of 100 photons per bit. Figure 9.19 shows examples of typical detector performances that have been achieved in practice.

The above discussion applies only to amplitude-modulated direct detection signals; other schemes have differing minimum signal requirements. Table 9.2 shows the number of photons per bit required for the coherent detection schemes discussed in sections 9.1.1 and 9.3.9 (see ref. 9.8).

At first sight the results in Table 9.2 seem rather puzzling, since it would seem that incoherent direct detection is almost the 'best' scheme available, and in view of the complexities involved with the coherent schemes (further discussed in section 9.3.9) it would seem rather pointless considering them at all. However, it has to be borne in mind that according

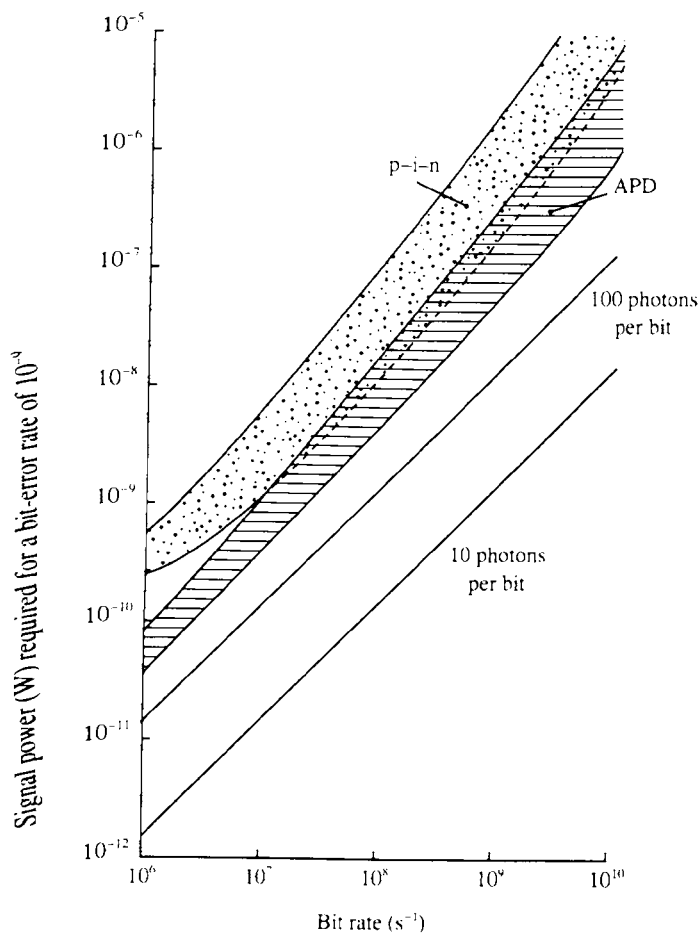


FIG. 9.19 Shaded regions indicate typical minimum signal powers required for p-i-n and APD receivers to achieve a bit-error rate of 10^{-9} as a function of bit rate at a wavelength of $1.3\ \mu\text{m}$. Also shown are the powers corresponding to a constant number of photons per bit. As demonstrated in the text, the line of 10 photons per bit represents the fundamental limit for this error rate.

TABLE 9.2 Photons per bit required to achieve a BER of 10^{-9}

	ASK	PSK	FSK
Incoherent	10		
Homodyne	18	9	
Heterodyne	36	18	18

to eq. (9.3) the local oscillator effectively boosts the photogenerated current by the factor $(P_{LO}/P_\lambda)^{1/2}$ where P_{LO} is the local oscillator power and P_λ is the signal power. This effective gain acts, as in the case of the APD, to cause the dominant noise factor at the output of the detector to be shot noise rather than Johnson noise. Thus for coherent systems the values in Table 9.2 represent much more realistically achievable values than is the case for incoherent detection.

9.3.3.4 Amplifier front end design

We have seen in the previous section that the characteristics of the first stage of the amplifier immediately following the detector can have a decisive effect on the ultimate performance of the detector/amplifier combination. In this section, we consider in more detail several ways in which this first stage (the amplifier 'front end') may be implemented.

We start with the so-called 'low impedance' front end. Referring to Fig. 9.17, the current source (i.e. the photogenerated current) 'sees' the diode capacitance (C_j) and load resistor (R_L) acting in parallel. With an amplifier connected across R_L , the bias resistor (R_a) associated with the amplifier first stage will also be in parallel with R_L . The effective load resistance (R_e) is then given by $R_e = R_L R_a / (R_L + R_a)$.

To obtain maximum signal bandwidth (see the arguments leading to eq. 7.35) both R_L and R_a should be as small as possible. The major penalty to be paid with such a design is that the system is then likely to be dominated by Johnson noise (see eq. 9.14), and the S/N ratio correspondingly poor. If the noise limitations are not a problem, however, then this design has the advantage of relative simplicity. The circuit is usually implemented using a bipolar junction transistor (see e.g. Fig. 9.20a, which shows a grounded emitter configuration).

If a low noise level is of paramount importance, then a 'high impedance' front end may be considered, that is R_a is made relatively large. The disadvantage is then that, since the electrical bandwidth is limited to $(2\pi R_e C_j)^{-1}$, the response will fall off at high frequencies as $1/f$, with the capacitor and parallel resistor combination acting to give an output proportional to the integral of the signal current. Thus it is only possible to have an extended frequency response by including a differentiating circuit (i.e. one whose output is proportional to f) later on in the amplifier chain. One problem with this latter approach is that, although it allows an increased bandwidth, it also restricts the available dynamic range, that is in the ratio of maximum to minimum usable input signals. A possible circuit for a high impedance front end is shown in Fig. 9.20(b). Here, a silicon field effect transistor (FET) is used in a common source configuration. The silicon FET exhibits both very high input impedance and low noise performance. One difficulty is that the gain of the device is rather limited: above 25 MHz the gain drops to values close to unity, and if modulation frequencies higher than this are required then a bipolar transistor is more often used.

A considerable effort has also been put into the development of GaAs FETs, since they can operate up to several gigahertz and thus offer an alternative to bipolar devices. There has been particular interest in hybrid integrated circuits utilizing p-i-n photodiodes with GaAs FET front ends. The integration technique enables stray capacitances to be reduced to very low values (e.g. approximately 0.5 pF).

A modification of the high impedance front end is the 'transimpedance' design. Here, a

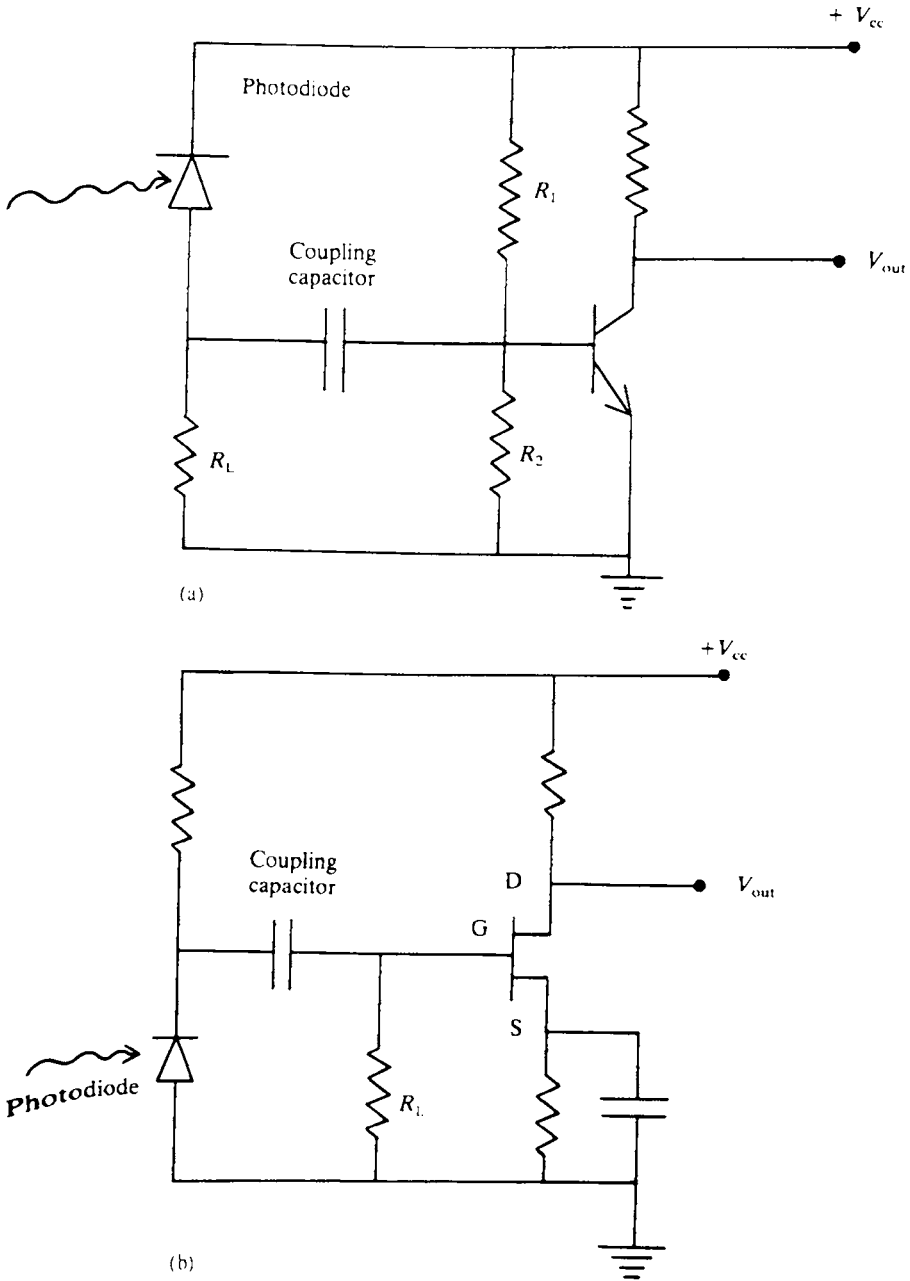


FIG. 9.20 Possible designs for front end circuitry: (a) a low impedance front end using a bipolar junction transistor (the value of R_1 is relatively low); (b) a high impedance front end using an Si FET (the value of R_1 is high).

high impedance amplifier is used in conjunction with negative feedback; that is, a portion of the output signal is fed back into the input with its phase reversed (Fig. 9.21). If the gain of the amplifier is G , then it is easy to show (Problem 9.13) that the magnitude of the output can be written

$$V_{\text{out}} = \frac{i_{\lambda} R_f}{(1 + 4\pi^2 f^2 C_j^2 R_f^2 / G^2)^{1/2}} \quad (9.21)$$

At relatively low frequencies the output of the detector is given by $i_{\lambda} R_f$ where i_{λ} is the photogenerated current and R_f the value of the feedback resistor. At higher frequencies the output will decline with increasing frequency, the bandwidth being given by $G/(2\pi C_j R_f)$ where C_j is the photodiode junction capacitance. As far as Johnson noise is concerned then it can be shown (ref. 9.9) that the effective noise resistance is approximately R_f . Thus if R_f is large the circuit will exhibit low Johnson noise; at the same time, provided that G remains high at high frequencies, the frequency bandwidth will also be relatively large thereby reducing or even eliminating the need for any differentiating circuitry.

9.3.3.5 Detector materials

The earliest detectors to be developed for use in fiber communication systems were based on silicon, which limited the operating wavelength to a maximum of $1.1 \mu\text{m}$. As we have seen in section 9.3.1, it is advantageous from the point of view of both fiber loss and material dispersion to use wavelengths around $1.3 \mu\text{m}$ or $1.55 \mu\text{m}$. Germanium detectors, which respond out to $1.9 \mu\text{m}$, are possible candidates. Unfortunately these suffer from much higher dark currents than do the corresponding silicon devices. In addition, germanium has a carrier ionization ratio, r , of about unity. This implies that the optimum gains in APDs are relatively

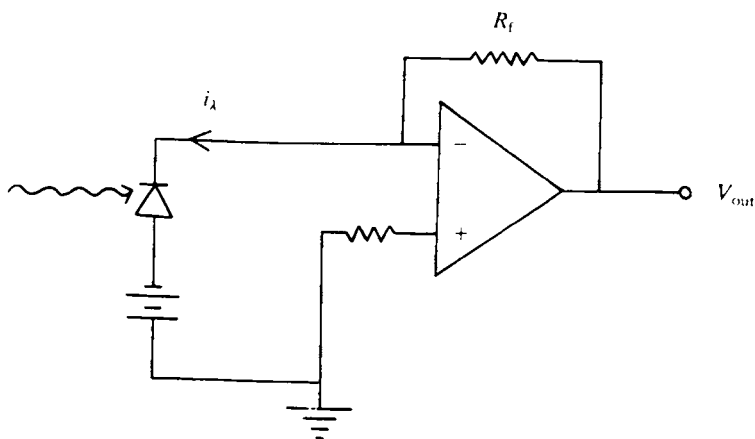


FIG. 9.21 Design for a transimpedance front end using an FET operational amplifier. At low modulation frequencies the voltage output is given by $V_{\text{out}} = i_{\lambda} R_f$, where R_f is the feedback resistor and i_{λ} the photogenerated current.

small, and hence also any improvements in S/N ratio. For example, taking $r = 1$ in eq. (7.38a) in conjunction with eq. (9.19) gives an optimum gain of 15 (see Problem 9.11). This implies an improvement in the optical power S/N ratio of only 8.9 (i.e. 9.5 dB).

As discussed in section 7.3.6.3 various types of detector can also be made from materials based on GaInAs and GaInAsP which have proved eminently successful for use in the 1.3 μm to 1.55 μm wavelength region.

9.3.4 Fiber choice

As we have seen in Chapter 8, there are three main types of fiber available, namely step index multimode, graded index multimode and single mode fibers. Step index multimode fibers are relatively inexpensive, have large NA values, but suffer from intermodal dispersion. Graded index fibers show greatly reduced intermodal dispersion but have relatively small NA values, and can carry only about one-half of the energy that can be carried by a step index fiber with the same core diameter. By their very nature single mode fibers are not subject to intermodal dispersion, instead their bandwidth performance is limited by material dispersion. They have such small diameters that usually only laser sources can couple sufficient amounts of power into them; in addition, a much higher accuracy is required in alignment at splices to avoid excessive jointing loss.

For many applications, a key consideration is the maximum length possible between transmitter and receiver (or between repeaters). Two factors influence this, namely fiber attenuation and fiber dispersion. In low bandwidth systems, the former is usually the limiting factor, whereas in high bandwidth systems it is more likely to be the latter. For example, suppose we consider a fiber with an attenuation of 5 dB km⁻¹; if a signal loss of some 40 dB can be tolerated, then a cable length of 40/5 or 8 km can be used. If the cable were of the step index type, however, with a bandwidth for 1 km of 20 MHz (i.e. a bandwidth-distance product of 20 MHz km), and a signal bandwidth of 10 MHz were required, then the maximum fiber length would be restricted to 20/10, that is 2 km. Figure 9.22 illustrates the effects of these factors on the maximum fiber lengths achievable for several fiber loss and dispersion values.

9.3.5 Optical amplifiers

At the start of section 9.3 we mentioned that over relatively long fiber links one or more repeater units may be required periodically to boost the signal. Such units are relatively complicated in their electrical circuitry requirements and can constitute a considerable portion of the cost of a long distance route. A much more elegant solution would be to develop an 'all optical' amplifier that did not require the optical signal to be turned into an electrical signal and then back into an optical one. The process of stimulated emission provides a mechanism whereby this may be achieved, since, bereft of its mirrors, a laser is simply an optical amplifier. In section 5.10.1 the fiber laser was briefly mentioned and is easily adapted to become an optical fiber amplifier. It is obviously important that such amplifiers operate at wavelengths of 1.3 μm and 1.5 μm to be compatible with existing optical fiber communication systems. Although operation at 1.3 μm has proved somewhat problematical, amplifiers

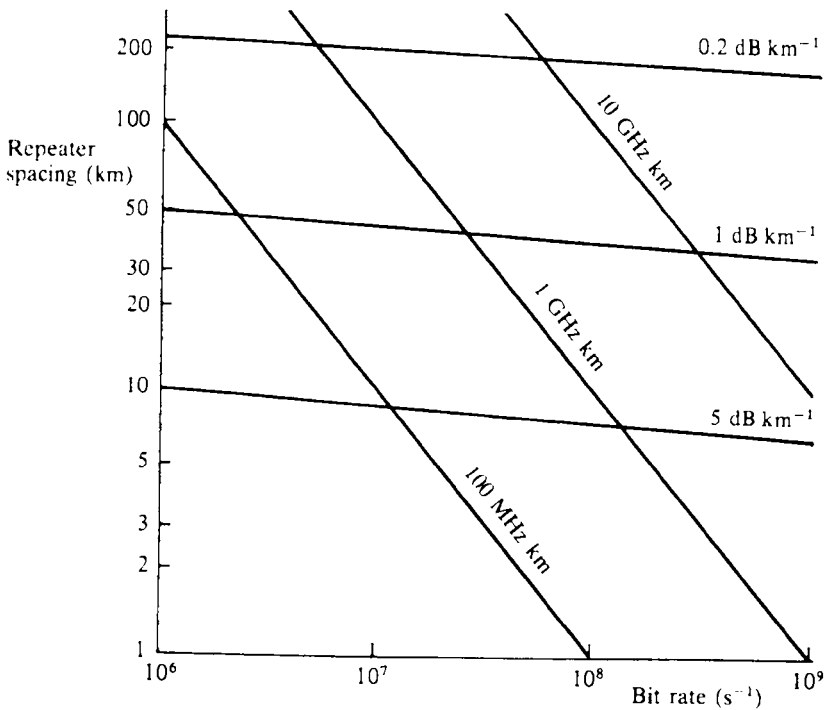


FIG. 9.22 Maximum repeater spacing allowing for differing fiber attenuations and dispersions as a function of bit rate. A total loss of 50 dB at 1 Gbps has been assumed. The attenuation-limited lines are slightly inclined owing to reduced receiver sensitivity with increasing bit rate.

based on erbium-doped fibers, which exhibit gain over a range of wavelengths centred on 1.55 μm , have been developed successfully. A key factor in this was the discovery that it was possible to incorporate relatively high concentrations of erbium into the fiber core (i.e. up to 1000 parts per million) when co-doping with alumina (Al_2O_3) is carried out.

The energy level diagram for the Er^{3+} ion in silica is indicated in Fig. 9.23. The first excited state ($^4\text{I}_{13/2}$) has a relatively long lifetime and gain is readily obtained once sufficient population inversion has been achieved between it and the ground state ($^4\text{I}_{15/2}$). (The system thus behaves as a ‘three-level system’.) Because the ions are incorporated into an amorphous solid matrix the energy levels are considerably broadened (section 5.10.1) and gain may be obtained over a range of some 30–40 nm. One big advantage is that, unlike the repeater unit, the operation of the device is independent of both the modulation coding format and (at least up to 100 Gbps) of the bit rate.

Many of the excited state levels can be used for pumping the laser, including the $^4\text{I}_{13/2}$ state itself.¹ For the present application, it is highly desirable that the pump source is compact, highly efficient and capable of coupling optical power into the fiber with minimal loss. All these requirements indicate the use of semiconductor lasers which limits the pumping wavelengths that can be used to 807 nm, 980 nm and 1.480 μm . Of these the 980 nm is the most effective, with small signal gains as high as 45 dB and incremental gains of 10 dB mW^{-1} of pump

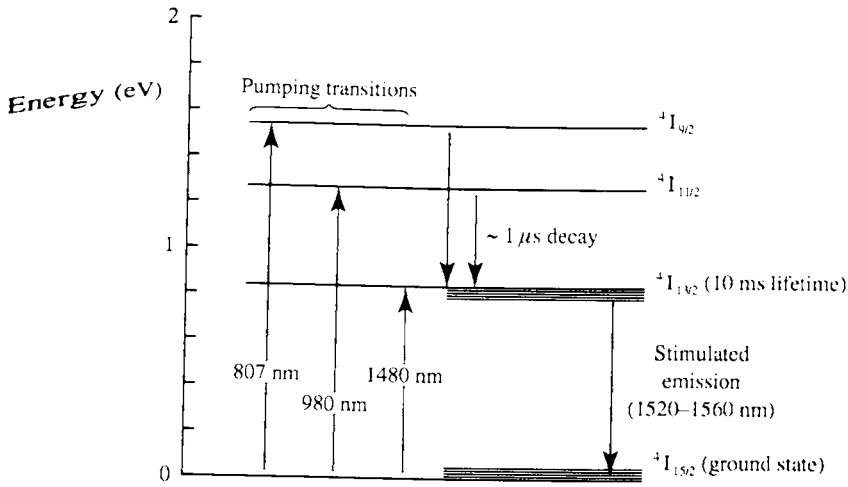


FIG. 9.23 The energy levels for Er^{3+} -doped silica fiber which are involved in the amplification of light over the wavelength range 1520 nm to 1560 nm using stimulated emission.

power having been achieved. Total output powers of up to 500 mW are also possible. By comparison the 807 nm pump wavelength suffers because atoms in the $^4I_{13/2}$ excited state can be excited by the pump wavelength to a yet higher state ($^2H_{11/2}$) thus reducing the amount of population inversion available. To generate a reasonably uniform gain across the fiber profile it is desirable that the fiber be single mode at the pump wavelength. This is obviously easier to achieve when the pump wavelength is 1.48 μm rather than 980 nm or shorter wavelengths. The efficiency is found to increase when the doping ions are concentrated in the centre of the core rather than being uniformly distributed.

Figure 9.24 illustrates a typical gain curve of a length of Er^{3+} -doped silica fiber. Ideally the gain curve should be flat so that all signals within the gain profile will be amplified by the same amount; in this case this is clearly not so. The degree of uniformity in the gain is affected by

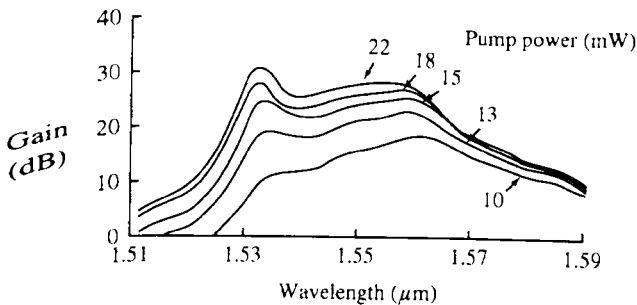


FIG. 9.24 Variation of gain with wavelength for a typical erbium-doped fiber amplifier. The results at several different pump powers with a wavelength of 1485 nm are illustrated.

a number of parameters including the concentrations of the active ion. It is possible to achieve a reasonably flat gain profile (or even one which allows for the variation in fiber attenuation with wavelength) by inserting a suitably designed optical filter into the system.

The physical layout of an erbium fiber amplifier is shown in Fig. 9.25. A means has to be found to inject both the signal and the pumping radiation into the core of the doped fiber section. This is normally achieved by means of a biconical fused fiber coupler (section 8.5.2). In this device the fractional amount of radiation coupled from one input fiber to the 'other' output fiber varies with the coupling length z as $\sin^2(Cz)$ (eq. 8.43). Since the value of C depends on wavelength and since pump and signal are at different wavelengths, it is possible to have a situation where the two different input fibers both couple with nearly 100% efficiency into the same output fiber as illustrated in Fig. 9.26 (this is not such an easy matter when both pump and signal have nearly the same wavelength, as, for example, when using the $1.45\text{ }\mu\text{m}$ pumping wavelength). When high output powers are required two pumps can be used, one at each end of the fiber. One of the pump beams will then be propagating along with the signal beam whilst the other will be propagating against the signal. This arrangement ensures that the population inversion, and hence gain seen by the signal, remains approximately constant along the fiber. It is usually necessary to prevent the pump radiation from entering the main fiber after the amplifier (or the section before the amplifier if a counter-propagating beam is also being used). It is also vital that optical feedback into the gain section from reflections outside be eliminated, since such feedback can give rise to unwanted laser action. Consequently polarization-insensitive optical isolators are often placed at both ends of the gain fiber. It should also be noted that in the absence of optical pumping the 'amplifier' will become highly absorbing so that backup pumping lasers should always be on hand in the event that the original pump lasers fail.

As far as amplification at a wavelength of $1.3\text{ }\mu\text{m}$ is involved the situation is not so clear cut. It is possible to use Nd^{3+} -doped silica core fibers but efficiencies are low because of

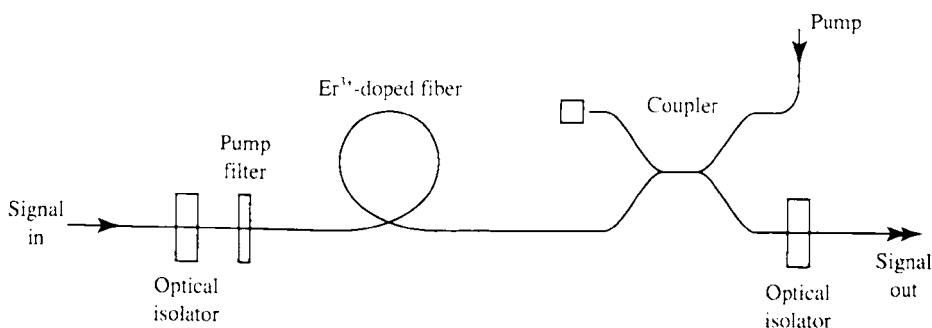


FIG. 9.25 The physical layout of an erbium-doped fiber amplifier. In this case the pump and signal beam are moving in opposite directions. A filter is inserted to prevent the small amount of pump radiation remaining after the gain section from propagating further along the fiber. In addition there are optical isolators at either end of the amplifier; these are to prevent any power at the signal wavelength from being reflected back into the gain section which could result in the device acting as a laser. More powerful pumping can be achieved with two pump beams moving in opposite directions.

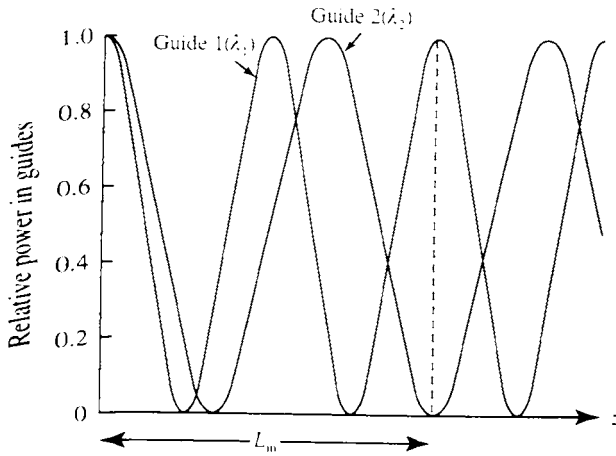


FIG. 9.26 Illustration of how a fiber coupler may be used as a mixer for two different wavelengths λ_1 and λ_2 . We assume that at the input to the mixer the power at wavelength λ_1 is guide 1 whilst guide 2 contains wavelength λ_2 . The amount of optical power remaining in the initial guide is shown as a function of coupling distance for the two wavelengths. After a distance L_m the power at wavelength λ_1 has transferred completely from guide 1 to 2, whilst that at wavelength λ_2 is all in guide 2. Thus if the coupler is terminated at this point the two inputs will be mixed together.

the presence of excited state absorption and competing radiative transitions. Praseodymium has also been tried as a dopant but appears to function best in host glasses such as 'ZBLAN' ($\text{ZrF}_4\text{--BaF}_2\text{--LaF}_3\text{--AlF}_3\text{--NaF}$) that are difficult to splice onto silica fibers.

Although the Er^{3+} -doped fiber amplifier has proved very successful there are other systems that have been proposed for use as optical amplifiers; one of the most important of these is the semiconductor laser amplifier. If the end facets of the semiconductor laser are suitably coated to reduce reflection then the device will function as an amplifier. The advantages of the semiconductor laser amplifier are a much greater freedom in the choice of operating wavelengths, which can be varied by varying the composition (and hence bandgap) of the semiconductor, and a broader and smoother gain curve. The main disadvantages are that the gain is dependent on the direction of polarization and that coupling losses between fibers and the amplifier can be significant unless great care is taken. It is possible that semiconductor laser amplifiers will find greater use in integrated optical circuits as discussed in section 9.4.

9.3.6 System design considerations

In this section we briefly consider typical component combinations used in direct detection systems. Three main factors are paramount in influencing system choice, namely bandwidth, maximum transmission distance and, of course, total system cost.

For short distances (i.e. up to a few hundred metres) and for bandwidths of up to a few tens of megahertz a cost-effective combination is a red-emitting LED, a high NA all-plastic fiber and an Si photodiode detector. More exacting distance and bandwidth requirements may

be met with a move to silica-based fibers with an LED (or laser) source operating at about 850 nm. The detector may be either a silicon photodiode or a silicon avalanche photodiode.

When the transmission rate exceeds 400 Mbps or so, the usable fiber length is determined primarily by fiber dispersion rather than by fiber attenuation (see Fig. 9.22). It then becomes necessary to move to single mode fibers and operating wavelengths close to the material dispersion minimum (i.e. near 1.3 μm in silica-based fibers). Transmission rates of several gigabits per second over fibers of several tens if not hundreds of kilometres in length have already been demonstrated or are being built.

The development of the erbium optical fiber amplifiers has made 1.55 μm an attractive wavelength to work at, with the added advantage that fiber losses (in dB km^{-1}) are about a factor of two smaller than at a wavelength of 1.3 μm . The downside to this is, of course, that material dispersion will be larger. There are two ways round this: first, if the installation is a new one then dispersion-shifted (or dispersion-flattened) fiber can be used. Secondly if a 'standard' fiber with a dispersion minimum at 1.3 μm is involved then a *dispersion compensation* scheme may be employed. In this a length of fiber which exhibits a high dispersion *in the opposite sense* to that of the main fiber is inserted just before the receiver. Provided the length has been chosen correctly the dispersion in the main fiber may be cancelled out by the dispersion in the opposite sense introduced by the compensation fiber to produce a 'dispersionless' link. Because the link now involves a longer fiber the total losses will be increased, but this increase need not be excessive since the fiber is operating at the wavelength of lowest loss.

To determine whether or not a chosen system will perform satisfactorily, a number of checks must be made. Obviously both detector and emitter must be capable of handling the required signal bandwidth. The fiber dispersion over the length required must not degrade the signal excessively, while for a given BER (for a digital signal) there will be a minimum average signal power that must reach the detector. If the power launched into the fiber is known, together with the fiber attenuation, the maximum length of fiber that can be used may be calculated. Allowance must be made for any splices or joins, and a safety margin of, say, 5 dB is also usually included. Such a calculation is called a *flux budget* (see Example 9.2). It is customary to use a logarithmic power unit called the *dBm*, where the power is referred to 1 mW; thus a power of 10 mW becomes $10 \log_{10}(10/1)$ or 10 dBm.

EXAMPLE 9.2 Flux budget calculation

Suppose we have to construct a link of length 15 km and bandwidth 100 Mbps. Components are chosen with the following characteristics: receiver sensitivity -50 dBm (at 100 Mbps), fiber loss 2 dB km^{-1} and transmitter launch power (into fiber) 0 dBm. It is anticipated that in addition 10 splices will be required, each involving a loss of 0.5 dB. The following table itemizes each loss term and allows a ready determination of whether the system should operate with sufficient power margin:

Transmitter output	0 dBm
Receiver sensitivity	-50 dBm
Required margin	50 dBm

System loss:

Fiber loss 2 dB km^{-1} , 15 km	30 dB
Detector coupling loss	1 dB
Total splicing loss ($0.5 \text{ dB} \times 10$)	5 dB
Headroom for temperature range, ageing effects and future splices	5 dB

Total attenuation 41 dB

Hence, the excess power margin is $50 - 41 = 9 \text{ dB}$, which should be sufficient for the operation of the link.

In Fig. 9.27 we show typical launch powers for both LED and laser sources and also the received powers required by state-of-the-art receivers to achieve a 10^{-9} BER as a function of bit rate. It is obvious from Fig. 9.27 that as the required bit rate increases the relative advantage of the laser/APD combination also increases.

Some indication of the overall performance of an optical fiber communication system can be obtained from the so-called 'eye diagram' (Fig. 9.28b). This is the pattern obtained on an oscilloscope when the sweep rate of the oscilloscope is set at a fraction of the bit rate, and the signal consists of a random sequence of ones and zeros. In an ideal system, where the received pulses were almost completely rectangular, the pattern would resemble Fig. 9.28(a), giving an open eye pattern. However, the received pulses will be broadened by dispersion/limited

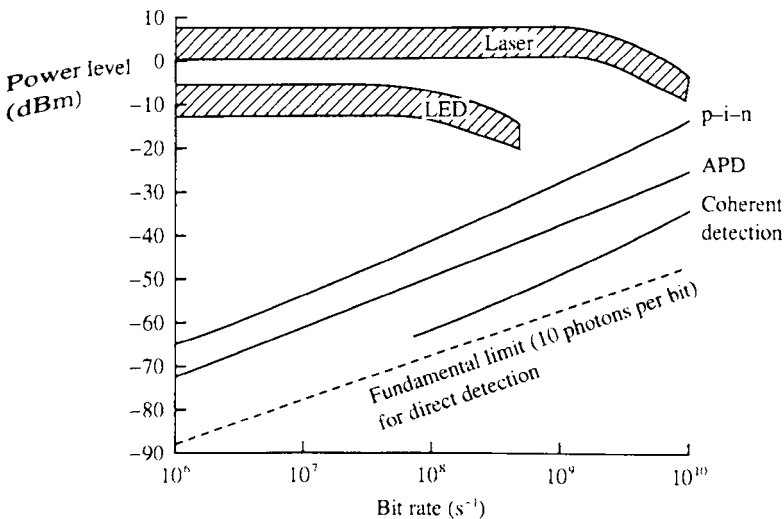


FIG. 9.27 Simple illustration of the system flux budget for representative systems. The upper shaded areas show typical launch powers from LED and laser sources, whilst the lower curves show typical receiver sensitivities for a 10^{-9} bit-error rate at $1.3 \mu\text{m}$ wavelength. The curves for the p-i-n and APD detectors are for direct detection; a further curve shows results for coherent detection. The fundamental detection limit corresponding to 10 photons per bit is also shown.

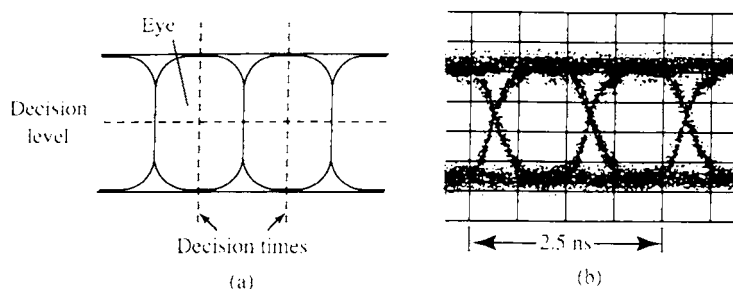


FIG. 9.28 (a) An idealized eye diagram where the pulses exhibit little dispersion and jitter. The decision times and decision levels are shown. (b) A typical eye diagram as obtained in practice.

receiver bandwidth and also subsequent pulses will be subject to a slight random displacement in time ('jitter'), and thus the eye will be 'closed' to some extent (as in Fig. 9.28b). The decision time and threshold level will be set with relation to the centre of the eye and it is evident therefore that the extent of eye closure will give an indication of the error rate.

So far in this section we have not discussed the use of erbium-doped fiber amplifiers (EDFAs). Over long distances they (together with repeater units) can be used to boost the signal periodically so that on arrival at the detector it is strong enough to give an acceptable BER (or S/N ratio). Optical amplifiers can be sited at three different positions in the optical fiber link to serve different purposes. First they can be placed immediately before the first section of optical fiber where they act as power amplifiers, the advantage being that only a low power source laser is required. Although direct modulation of high power lasers is possible they are prone to exhibiting a wavelength 'chirp', that is a change in emission frequency throughout the emitted pulse as described at the end of section 9.3.2. This has the effect of increasing the effective source linewidth and so limiting the system bandwidth because of increased material dispersion.

Another position for the amplifier is just before the receiver. This boosts the incoming signal and provides increased signal sensitivity; here only a low signal power is involved, in contrast to the previous usage. It is of interest to note that a fiber amplifier p-i-n receiver combination can achieve sensitivities which almost equal those offered by coherent detection systems.

Finally the optical amplifier can take the place of the repeater unit. It should be remembered that the optical amplifier can only amplify, it cannot reshape or retime the optical pulses. Furthermore the amplifiers do not provide noise-free amplification. Whenever the amplifier is being pumped there will be some spontaneous emission generated, most of which will lie within the gain bandwidth. Some of this radiation will propagate down the fiber along with the signal and will be amplified at each subsequent amplifier. This will degrade the system performance in two ways: first the S/N ratio will worsen after each amplifier, and secondly the amplified spontaneous emission will serve to saturate amplifiers further down the chain thus limiting the available gain.

Finally in this section, to illustrate some of the rapid advances in technology that have taken place in optical fiber communication links, we look briefly at three transatlantic

telecommunication links. TAT 8 was the first ever optical fiber transatlantic link and was installed in 1988. Details are provided in Table 9.3. Also shown are details of TAT 9 and TAT 12. The latter, when introduced, provided more capacity than all of the previous links put together. It was also one of the first major systems to employ erbium-doped fiber amplifiers (EDFAs) in place of repeaters. All three systems cover a distance of about 6000 km.

9.3.7 Local area networks

Although the initial drive to develop optical fibers has come mainly from the telecommunications area, that is from the need to develop high bandwidth, long distance, repeaterless point-to-point links, there are other types of link, for example the local area network (LAN), to which optical fibers may make significant contributions. The LAN concept is difficult to specify exactly, but aims to provide a communication network between a group of users within a restricted geographical area (e.g. within a factory or group of offices). Typical maximum distances involved might be a few kilometres. Local networks involving telephone links are nothing new, of course, but the availability of much increased bandwidths enables several other types of communication to be contemplated, for example data communication and even video links. Situations can easily be envisaged where networks linking up to, say, 100 users with data rates of up to several megabits per second may be required.

One of the main problems with realizing such networks using optical fibers is that the most loss the system can readily tolerate between launch and detection is about 40–50 dB (i.e. see Fig. 9.27). With coaxial cable, the corresponding figure is nearly 90 dB. In long haul communication links, the advantage this gives to coaxial links is more than offset by their relatively high transmission losses. For the comparatively short distances involved in typical LANs, however, coaxial links do not have this built-in disadvantage. Furthermore, any such system will inevitably involve a large number of 'taps' and switches. For an optical system these components, at present, involve appreciable inherent losses, whilst the corresponding losses in copper wire technology are negligible. Nevertheless, fiber systems do offer the advantages of high bandwidth, immunity from local interference, a higher level of information security and small size and weight of the cable.

There are many ways in which the interconnections in LANs may be made with a trade-off occurring between such factors as system complexity (and hence cost) and flexibility. Figure 9.29 illustrates several possible topologies for such systems. In Fig. 9.29(a), the 'fully connected' network, every potential user (or 'node') is connected to every other. This arrangement is complex, expensive and needs considerable rewiring if more nodes are added. Some of the links may never be used. It is, however, the most reliable topology, since if one link

TABLE 9.3 Some recent transatlantic telecommunications systems

	Year	Wavelength (μm)	Bit rate	Repeater/EDFA spacing
TAT8	1988	1.3	280 Mbps	50 km (repeater)
TAT9	1989	1.55	565 Mbps	100 km (repeater)
TAT12	1995	1.55	5 Gbps	45 km (EDFA)

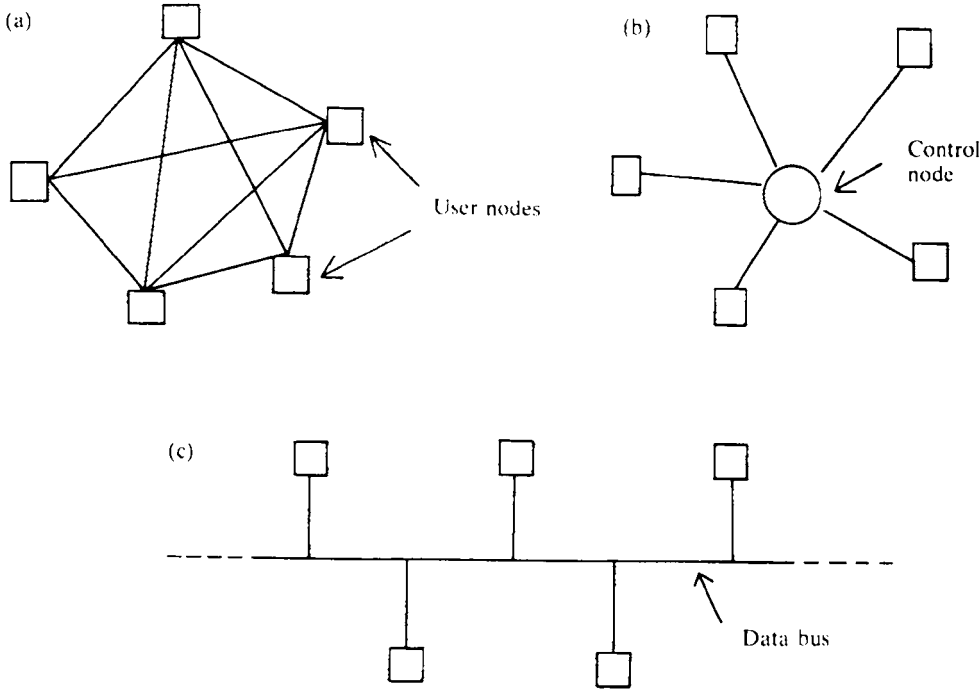


FIG. 9.29 Three possible topologies for a local area network: (a) fully connected, (b) star, and (c) data bus.

is damaged then only this one linkage is affected, and rerouting may be possible to restore the connection.

Figure 9.29(b) illustrates the 'star' network. Here, each node is connected to a central 'control' node. If this is 'active', each signal can be switched to its required destination. If 'passive', a particular node will direct a signal to all nodes, each of which will then have to decide themselves whether or not to accept the signal. The number of links is reduced to a minimum, but the reliability of the whole system hinges on that of the central node. If this fails in some way, the entire system is likely to break down. Figure 9.29(c) illustrates the 'data bus'; this offers great flexibility and is efficient in its use of fiber.

Some of the optical components likely to be useful in LANs have been touched upon in section 8.5.2. For example, the 'mixing rod' type of fiber coupler could form the basis of the central (passive) node of a star network. Similarly, the fused biconical coupler could be designed to provide a small amount of coupling from one fiber into another, thus providing a 'tap' as required in a data bus configuration.

The diversity of possible systems and their realization make any comprehensive discussion here impossible, and the interested reader may consult the texts of ref. 9.10 for further information.

9.3.8 Wavelength division multiplexing

One way of increasing the bandwidth of an existing optical link with no modification to the

fiber itself is to employ the technique of *wavelength division multiplexing* (also known as *optical frequency division multiplexing*). In this, signals on different carrier wavelengths are mixed together (i.e. multiplexed) and then transmitted simultaneously down a fiber. This technique has the obvious advantage of increasing the capacity of an optical fiber system without changing the fiber. It can also play an important role in networks. For example, consider a number of users connected to a central hub in the star network configuration of Fig. 9.29(b). Each user is assigned a particular transmission wavelength and the hub site transmits all the received messages to the other sites who can then decide which user they wish to be connected to.

The signal multiplexing may be carried out using some type of mixer such as the 3 dB coupler (as discussed in section 8.5.2), although there will be significant losses associated with this. A somewhat more difficult task is the wavelength separation at the detector end. Again couplers may be used to split the signal up amongst the number of receivers required. The individual wavelengths required by each receiver can then be isolated using some type of narrow optical bandpass filter (usually of the multilayer dielectric type). Again there are inevitable losses which become increasingly large when the number of separate wavelength channels increases. Typical channel separations of 4 nm with cross-talk of better than 30 dB may be obtained at 1.3 μm wavelength. The recent development of Bragg fiber gratings structures (section 8.7.5.2) has opened up the possibility of in-fiber wavelength-selective devices. For example, the wavelength notch filter required for wavelength demultiplexing can be obtained by using Bragg reflection gratings with staggered wavelengths such that there is a transmission gap in their resonant spectra.

An alternative wavelength demultiplexing technique is to use some type of spatial dispersion element such as a prism or (more usually) a diffraction grating. A schematic diagram of such a system is shown in Fig. 9.30. Light from a single fiber is collimated with a lens and then dispersed spatially using a blazed diffraction grating. The lens refocuses each individual wavelength onto its own output fiber each of which is connected to an individual

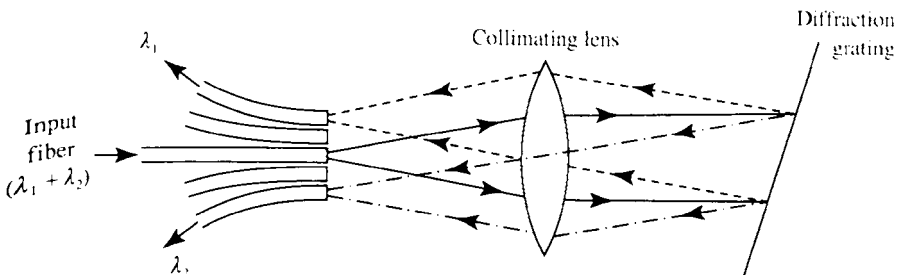


FIG. 9.30 Illustrating one method that can be used to separate a wavelength-multiplexed signal into its two (or more) constituent wavelength signals. Light (here containing two wavelengths, say λ_1 and λ_2) from the input fiber is collimated, and then falls onto an inclined diffraction grating. This light is diffracted back but with different wavelengths being diffracted at different angles. The lens refocuses the light on the fibers with different wavelengths being focused onto the ends of different fibers. The different wavelengths are then taken to the appropriate detector for that particular channel.

detector. Up to about 20 channels can be accommodated with channel separations of down to 4 nm, insertion losses of about 5 dB and cross-talk between channels of 20 to 30 dB. More details concerning possible wavelength multiplexing/demultiplexing techniques are given in ref. 9.11.

Although the idea of wavelength division multiplexing is attractive its implementation is not straightforward, one notable problem being the inevitable power loss that is involved in the multiplexing/demultiplexing process. Thus each individual channel has to have a significantly higher power than in a 'normal' single channel system. Where there is quite a large number of channels involved the fibers may be carrying a significant amount of total power, so much so that non-linear processes such as stimulated Raman scattering and stimulated Brillouin scattering (see section 10.1.4) become a possibility. In addition if optical fiber amplifiers are used to boost signal levels the relatively large amounts of total power involved may cause the amplifiers to saturate and hence reduce the gain available. In addition the problems associated with amplified spontaneous emission will be more severe than with single wavelength systems. It should also be remembered that the gain is not flat over the gain profile of the amplifier. The usual way of dealing with this is to use some form of compensating filter; however, even small differences in gain can become significant after a number of amplifications have taken place.

9.3.9 Coherent systems

The basic theory behind the operation of coherent systems was discussed in section 9.1.1. The output signal from the fiber is mixed with that from a local oscillator (usually a semiconductor laser) and the combined fields are then incident on the detector. In the case of heterodyne detection, where the signal and local oscillator frequencies (ω_c and ω_o respectively) are different, the output of the detector is given by

$$O_d = RA_c A_o \cos[(\omega_c - \omega_o)t + (\phi_c - \phi_o)] \quad (9.3)$$

In the case of homodyne detection (where $\omega_c = \omega_o$) we have

$$O_d = RA_c A_o \cos(\phi_c - \phi_o) \quad (9.4)$$

Homodyne detection is difficult because of the problem of ensuring that both the signal and local oscillator frequencies remain equal, and it is rarely used in practice. Heterodyne detection is certainly practicable, however, and it has two main advantages over direct detection. First it can operate with signals which are close to the quantum noise limit and secondly, by varying the frequency of the local oscillator, the receiver can 'tune in' to any signal frequency of interest provided it is within the tuning range of the local oscillator. It thus provides a possible solution to the problem of separating individual signals from the group of signals involved when using wavelength division multiplexing. However, it has to be admitted that coherent detection involves a considerable increase in system complexity, and hence cost, over direct detection systems.

One of the obvious requirements is for an optical source (the local oscillator) with a highly stable output with respect to both amplitude and frequency. In practical terms the largest linewidths that can be used for both ASK and FSK are about 20 MHz and for PSK

about 1 MHz. The standard Fabry–Perot semiconductor laser has a linewidth of about 10 GHz (when directly modulated this may increase to 1000 GHz). It can be reduced in DFB lasers (see section 6.2) to between 1 and 100 MHz (i.e. suitable for ASK and FSK), and in large external cavity lasers to below 100 kHz (i.e. suitable for PSK). As far as tuning in to closely spaced signals is concerned field trials have indicated that it is possible to have channels separated by about 10 GHz. This is certainly superior to typical channel separations in normal wavelength division multiplexing schemes where the channels have to be usually at least 150 GHz apart (i.e. 1 nm apart in wavelength terms).

Another factor to be taken into account is that of the stability of the polarization of the signal within normal communications fibers. When considering coherent detection we have assumed that the polarization state of the local oscillator is exactly the same as that of the signal. Single mode fiber can support two orthogonally polarized modes and any outside disturbances to the fiber, such as temperature or stress changes, can give rise to a varying amount of coupling between these two (i.e. a type of mode coupling) and hence lead to unpredictable changes in the polarization state of the signal. This in turn will cause signal ‘fading’ unless measures are taken to counter the problem (if the signal and local oscillator have orthogonal polarizations then no signal at all will be received). The use of polarization-maintaining fiber is one obvious solution, though it is considerably more costly than standard fiber, and the fact that all the existing fiber links would have to be replaced make this solution a virtual non-starter. A more realistic approach is to track the state of polarization at the receiver and then to change the state of polarization of the local oscillator to match that of the signal. This is in fact a reasonable proposition since the changes in polarization take place relatively slowly (over timescales of the order of seconds) and the light from the local oscillator can have its polarization state changed by passing it through a length of fiber which is subject to variable asymmetric stress. Figure 9.31 illustrates the schematic layout of such a system. The two input signals are mixed together using a 3 dB coupler and the radiation from both output fibers is utilized in what is termed a *balanced receiver* configuration. It may be shown that the two outputs from the 3 dB coupler have a phase difference of π . If, therefore, we take the difference between the outputs of the two detectors by having them connected back to back as shown, we effectively double the signal that would be available from each detector individually.

Other techniques which can be used to deal with the polarization stability problem include *polarization diversity* and *polarization scrambling*. In the former the incoming radiation is split into two and each part falls onto a separate detector. The two detectors are connected in a balanced receiver configuration and are set up so as to respond to light that is polarized in orthogonal directions. Whatever the polarization of the incoming radiation one or both of the detectors will be able to respond. In the polarization scrambling technique the polarization of the source is deliberately varied so that it undergoes a complete rotation of the polarization during the bit period. Thus there will always be, at some time during the bit period, an output from the detector. Both these techniques have been demonstrated successfully, but the latter incurs a 3 dB power penalty, that is twice as many photons per bit are required to achieve the same BER as in the ideal coherent system.

At present the number of coherent systems in commercial use is insignificant. This is mainly because the much simpler erbium fiber amplifier + p–i–n direct detection systems

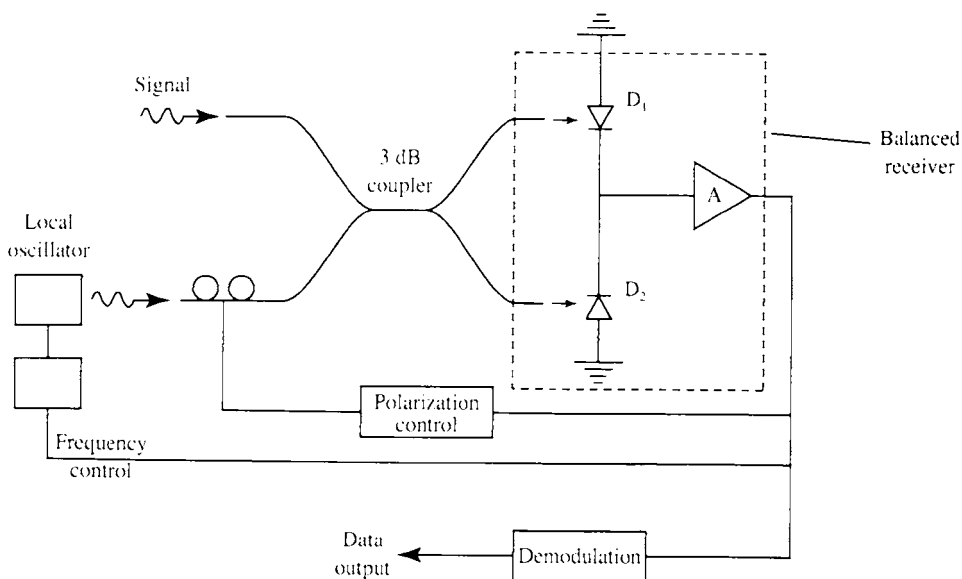


FIG. 9.31 Diagram showing the layout of a coherent receiver using a balanced receiver configuration.

are capable of detection sensitivities which are within a few decibels of those obtainable from coherent systems (see Fig. 9.27). This has removed one of the main attractions of coherent systems, though their superiority over direct detection in terms of channel spacing and tuning range has not yet been utilized.

9.3.10 Solitons

One of the fundamental limitations on bandwidth in single mode optical fiber systems is due to material dispersion (section 8.3.5). It arises because different wavelengths of light travel at different velocities in a dispersive medium. From the discussion leading up to eq. (8.34) we may write

$$\frac{d\tau}{d\lambda_0} = -\frac{L}{c} \lambda_0^2 \frac{d^2n}{d\lambda_0^2} \frac{1}{\lambda_0}$$

where τ is the time taken for radiation of wavelength λ_0 to traverse a distance L of fiber. At a wavelength which is greater than the minimum dispersion wavelength, the quantity $\lambda_0^2 d^2n/d\lambda_0^2$ is negative (see Fig. 8.26) and hence $d\tau/d\lambda_0$ will be positive. In this situation longer wavelengths will travel more slowly than shorter wavelengths and any optical pulse will spread out in time as it travels down the fiber with the shorter wavelengths being at the leading edge and the longer wavelengths at the trailing edge. However, there is another mechanism present which, under certain circumstances, is capable of opposing this pulse broadening. This arises because the electric field in the pulse causes a change in the local refractive index via the Kerr effect (eq. 3.16). As a consequence a phase shift develops at the leading

edge of the pulse which causes the radiation there to move to lower optical frequencies (i.e. a 'red shift' develops). Conversely at the trailing edge there is a corresponding blue shift. This wavelength variation through the pulse is in the opposite sense to that induced by material dispersion and hence there is the possibility that a non-dispersive pulse can be obtained, provided that the magnitudes of these two dispersive effects can be made equal. This can be shown to be possible provided that the pulse shape takes on a particular form, and also that the maximum pulse power exceeds a particular threshold value (which is dependent on both the loss within the fiber and the Kerr coefficient). Within a typical low loss silica fiber the minimum pulse power required is of the order of a few tens of milliwatts with corresponding pulse widths of a few tens of picoseconds. The required pulse shape is given by (ref. 9.12)

$$I(t) = I_0 \text{sech}^2(t/T_s) \quad (9.22)$$

where the parameter T_s varies inversely as the square root of the peak power. It may be noted that the total energy in such a pulse is $2I_0T_s$ whilst the pulse width is $1.76T_s$. Thus the pulse width becomes narrower as the total energy and the peak power increases. The non-dispersive pulse is known as a *soliton*.²

Once a soliton has formed it tends to be remarkably stable, passing splices etc. with only momentary disturbances and adapting its shape to suit the local conditions. However, because of fiber attenuation, as the pulse progresses it will gradually lose energy, and hence broaden (Fig. 9.32). When the pulse energy falls below the threshold value then the pulse loses its soliton characteristics. However, optical amplifiers (as described in section 9.3.8 above) can be used to restore the pulse energy when it nears the threshold value.

Transmission of soliton pulses at data rates of several gigabits per second over several thousands of kilometres of fiber has been demonstrated on a number of occasions (see e.g. ref. 9.13). So far soliton transmission remains a very interesting possibility; whether it will be used seriously for high bit rate and very long distance transmission remains to be seen, although it is an obvious method of overcoming the problems caused by material dispersion which is present in all fibers.

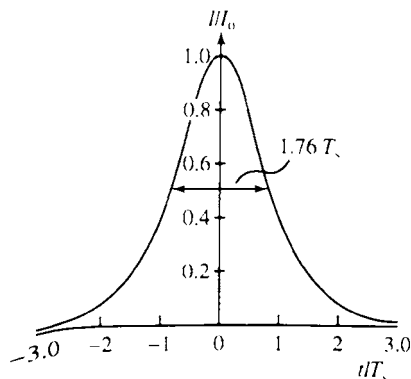


FIG. 9.32 The irradiance profile of a first-order soliton, with a sech^2 time dependence.

Although signal transmission using light waves is now well established, the optical signal usually has to be converted back into an electrical one if any processing needs to be done. The aim of integrated optics (IO) is to be able to carry out as much signal processing as possible on the optical signal itself. It is envisaged that a family of optical and electro-optical elements in thin film planar form will be used, allowing the assembly of a large number of such devices on a single substrate. Most device elements are expected to be based on single mode planar optical waveguides. Similar advantages are expected to those accruing when the idea of the integrated circuit was adopted in electronics.

The basic concept of IO was first proposed by Anderson in 1965 (ref. 9.14) and considerable progress has been made since then. (For a fairly comprehensive coverage of the field, the reader may consult ref. 9.15.) Initially most effort was put into demonstrating the viability of a wide range of individual devices, whilst latterly interest has concentrated on the problem of device integration. One of the main difficulties has been that no one substrate is ideally suited for all the different types of device. Many of the earlier composite devices were based on hybrid structures (see e.g. Fig. 9.48) but recently more fully integrated structures have begun to emerge.

9.4.1 Slab and stripe waveguides

It is generally assumed that in IO the signal will be carried within planar waveguides in either slab or stripe form which are formed by modifying the surface of a substrate. Planar waveguides were discussed in section 8.2, but there we confined our attention to symmetrical guides. In the present instance we are usually dealing with *asymmetrical* guides, that is guides where the layers above and below the guiding layer have different refractive indices (Fig. 9.33). The topmost layer (of refractive index n_0) is often air and consequently has a much lower refractive index than either the guide layer (n_1) or the substrate (n_2); such a guide is

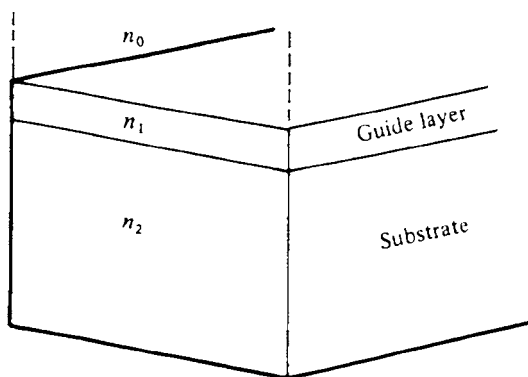


FIG. 9.33 Slab planar waveguide. The guide itself is formed on a substrate. The medium above the guide is usually air. The refractive indices of substrate, guide and topmost layer are n_2 , n_1 and n_0 respectively.

thus referred to as a *strongly* asymmetric guide. It is a relatively simple matter to extend the treatment of symmetric guides given in section 8.2 to cover the strongly asymmetric case.

We suppose that the phase changes at the upper and lower interfaces are $\phi_1(\theta)$ and $\phi_2(\theta)$ respectively, and using similar arguments that led up to eq. (8.9) the condition for a ray to be able to propagate becomes

$$\frac{4\pi n_1 d \cos \theta}{\lambda_0} - \phi_1(\theta) - \phi_2(\theta) = 2m\pi$$

For a ray to undergo total internal reflection at both the upper and lower interfaces, the internal guide angle must always be greater than either of the critical angles of the upper and lower interfaces (denoted by θ_{c1} and θ_{c2} respectively). Because we have a strongly asymmetric guide the critical angle at the upper interface will be much smaller than that at the lower. Thus as far as guided rays are concerned, we can assume that the internal guide angle will always be much larger than the critical angle of the lower interface, and consequently we may approximate $\phi_1(\theta)$ by π . The propagation condition can now be written

$$\frac{4\pi n_1 d \cos \theta}{\lambda_0} = \phi_2(\theta) + (2m + 1)\pi \quad (9.23)$$

Following a similar line of reasoning that led up to eq. (8.12), the condition for the $m = 1$ mode to propagate becomes

$$\frac{4\pi n_1 d \cos \theta_{c2}}{\lambda_0} > 3\pi$$

In contrast to the symmetric waveguide case, it is now possible for no mode at all to propagate, which happens when

$$\frac{4\pi n_1 d \cos \theta_{c2}}{\lambda_0} < \pi$$

Thus the condition for only a single mode to propagate now becomes

$$\pi \leq \frac{4\pi n_1 d \cos \theta_{c2}}{\lambda_0} \leq 3\pi$$

By putting $\cos \theta_{c2} = [1 - (n_2/n_1)^2]^{1/2}$ this condition can be written

$$\frac{\pi}{4} \leq V \leq \frac{3\pi}{4} \quad (9.24)$$

where V is as defined in eq. (8.14). In terms of the guide thickness d we have

$$\frac{1}{4(\text{NA})} \leq \frac{d}{\lambda_0} \leq \frac{3}{4(\text{NA})} \quad (9.25)$$

where $\text{NA} = (n_1^2 - n_2^2)^{1/2}$.

This equation implies that, in contrast to the symmetrical guide case, there is a minimum thickness below which it is not possible for the guide to support a single mode. A calculation based on this equation is given in Example 9.3.

EXAMPLE 9.3 Guide thicknesses for strongly asymmetric waveguides

Planar waveguides may be made in LiNbO_3 by diffusing in titanium; a 1% concentration of titanium causes (at a wavelength of $0.63\text{ }\mu\text{m}$) a refractive index change of about 6×10^{-3} . The ordinary refractive index of LiNbO_3 is 2.286, so we have $n_1 = 2.286$ and $n_2 = 2.280$ and then $\text{NA} = (2.286^2 - 2.280^2)^{1/2} = 0.166$. Using eq. (9.25) the requirement for single mode behaviour becomes

$$1.51 \leq d/\lambda_0 \leq 6.02$$

The exact value chosen for d/λ_0 depends on the manufacturing process and the use to which the guide is being put. Assuming that it is desirable that the guide be as thick as possible then a suitable design thickness for such a guide could be $d = 5\lambda_0$. It would be unwise to approach the upper limit too closely lest the inevitable fluctuations in d that will occur in any manufacturing process cause the guide to become multimode in some places.

The field distribution can also be derived using a similar approach to that used for the symmetric guide, and is illustrated for a TE_0 -type mode in Fig. 9.34. A cosinusoidal variation is still obtained across the core with the peak being displaced towards the n_2/n_1 interface. Similarly an exponential decline is observed in the cladding but this is now more rapid in the upper medium (n_0) than in the lower (n_2).

The above analysis assumes an infinitely wide waveguide; in practice most waveguides used in IO have an approximately rectangular cross-section so that there is confinement in both the x and y directions. Some typical waveguide configurations are illustrated in Fig. 9.35. A further complication is that the waveguides usually have graded index profiles. There are no general explicit solutions available for either of these situations. A very simplistic

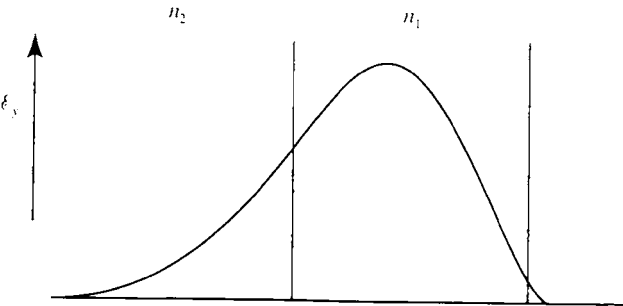


FIG. 9.34 The TE_0 field distribution within a strongly asymmetric planar waveguide, where $n_1 - n_0 \gg n_1 - n_2$.

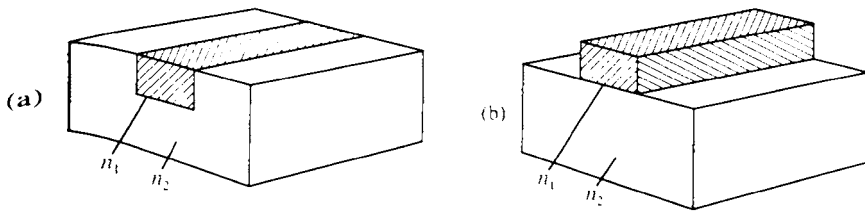


FIG. 9.35 Two basic geometries used for making IO stripe waveguides: (a) the channel waveguide; (b) the ridge waveguide.

view is to assume that, within the waveguide core, the modal field variation with x and y will be given by the product $\xi_x(x)\xi_y(y)$, where the field $\xi_x(x)$ is the solution to the waveguide equation given by assuming the guide is confined in the x direction but unconfined in the y direction (and vice versa for $\xi_y(y)$). Such solutions can be useful as the starting point for more accurate 'perturbation'-type calculations, but with the powerful computing facilities now widely available, the field distributions are more readily determined from 'first-principle' calculations involving the point-to-point solution of Maxwell's equations across the guide cross-section.

It is possible to make a few simple generalizations based on our understanding to date. We would expect that in single mode guides the dimensions in both the x and y directions should be of the order of a few times the propagating wavelength (see Example 9.3), and that the mode fields should peak within the core of the guides, declining towards the edges of the core and having a quasi-exponential decline with distance away from the core in the cladding regions.

The three main types of material which have been used as the basis of integrated optical types of waveguides are various types of glass, materials with high electro-optic coefficients such as lithium niobate, and semiconductor materials such as GaAs. A wide variety of different techniques (see ref. 9.16) can be used in waveguide manufacture: examples include sputtering one type of glass onto another, in-diffusion of a layer of titanium deposited on a substrate of lithium niobate, and liquid phase epitaxy. This last technique can be used with semiconducting materials such as GaAs and GaAlAs. The physical extent of the waveguide can often be delineated using the same photo- or electron beam lithography techniques that are common in the semiconductor integrated electronics industry. Losses in IO waveguides are usually much higher than in optical fibers, being of the order of 0.1 dB mm^{-1} . One reason for this is the large scattering from the upper waveguide surface (i.e. the waveguide/air interface) which is often relatively rough. Another problem, which may hinder miniaturization, is that any bends in the waveguides with small radii of curvature (i.e. less than a millimetre or so) must be avoided, otherwise losses can become prohibitively large (see the discussion in section 8.4.1).

9.4.2 Basic IO structural elements

We consider first of all one of the simplest passive devices, a waveguide splitter where an initial single mode waveguide splits into two single mode waveguides of the same width

(Fig. 9.36). In the first section of this device the waveguide expands gradually to twice its original width of a and then the guide splits into two sections each of width a and angled at θ to the original direction. Energy that is in the initial section of the guide will divide itself equally between the two output guides. The loss at the splitting point can be estimated by calculating the overlap integral (eq. 9.9) of the mode field in the section of width $2a$ with the fields in either of the two output sections which are angled at θ to it. An example of such a calculation is given in ref. 9.17; the outcome is that to avoid excessive loss the angle between output waveguides ($= 2\theta$) must be small (usually 1° or less). It should be noted that in this and several subsequent diagrams involving waveguide splitters the angles are shown as being much larger than 0.5° ; this is merely for pictorial convenience.

One of the simplest active devices is a phase modulator which is similar to the Pockels cell discussed in section 3.4. A stripe waveguide is formed within a suitable optically active material such as lithium niobate and electrodes are formed on the substrate surface on either side of the guide (Fig. 9.37). If the electrodes are a distance D apart and extend for a length L , then the additional phase shift $\Delta\phi$ produced when a voltage V is applied across the electrodes is, from eq. (3.14),

$$\Delta\phi = \frac{\pi}{\lambda_0} r n_i^3 V \frac{L}{D}$$

(9.26)

where r is the guide material electro-optic coefficient (see Table 3.1 for representative values).

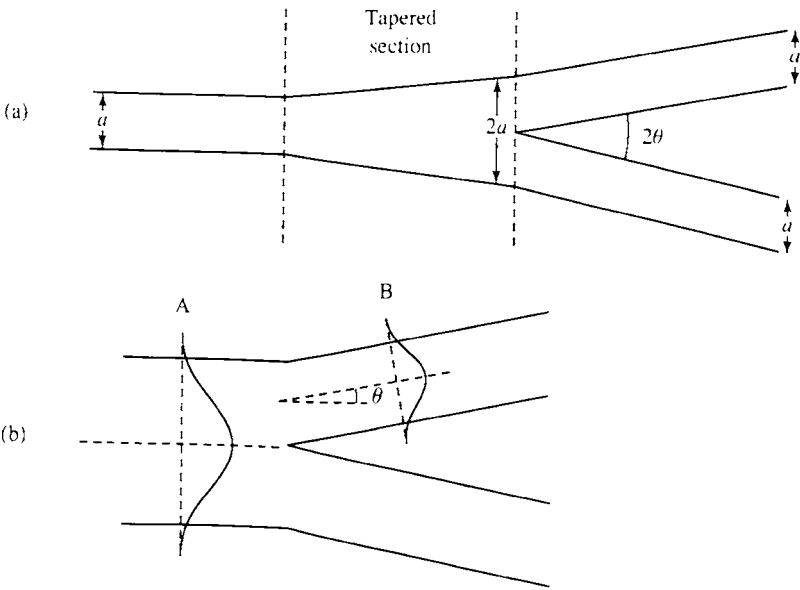


FIG. 9.36 (a) A waveguide 'Y' branch with an angle 2θ between the guides. Before the actual branch point there is a tapered section where the waveguide width is slowly increased from a to $2a$. (b) The efficiency of the splitting process as a function of θ may be calculated by determining the overlap integral of the waveguide modes at points just before (A) and just after (B) the branch point.

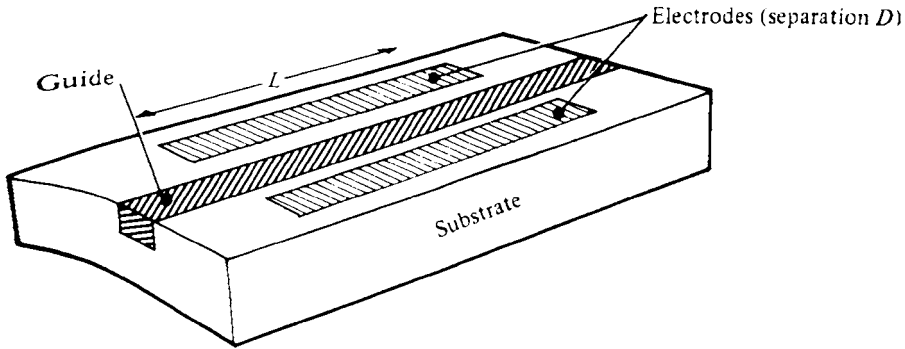


FIG. 9.37 Integrated optical version of the Pockels cell that may be used as a phase shifter.

A big advantage over bulk Pockels effect devices is that the ratio L/D may be made relatively large, say about 1000, and a phase change of π may then be achieved with voltages as low as 1 V or so (see Problem 9.15).

A high speed switch/modulator may be made by incorporating the phase shifter into one arm of the interferometric arrangement shown in Fig. 9.38. This configuration is known as a *Mach-Zehnder* interferometer. In it the guide splits into two with both paths rejoining after an identical path length. With no applied voltage across the phase shifter, the radiation in the two arms will have the same phase when they recombine, and hence the device will not affect the radiation flowing along the guide.³ However, if the phase shifter is activated to give a phase shift of π , then, on recombining, the radiation in the two arms will destructively interfere and no radiation will proceed down the guide (see Problem 9.15). It is easy to show that the output of the device will have a sinusoidal dependence on the applied field (see section 10.1.3.1 which deals with a very similar problem), and so if the device is biased at the half maximum transmission point the output irradiance will vary approximately linearly with applied field provided the field variations are small (see Fig. 3.10b for an illustration of this). Devices are commercially available, based mainly on lithium niobate, which have modulation capabilities of up to a few tens of gigahertz.

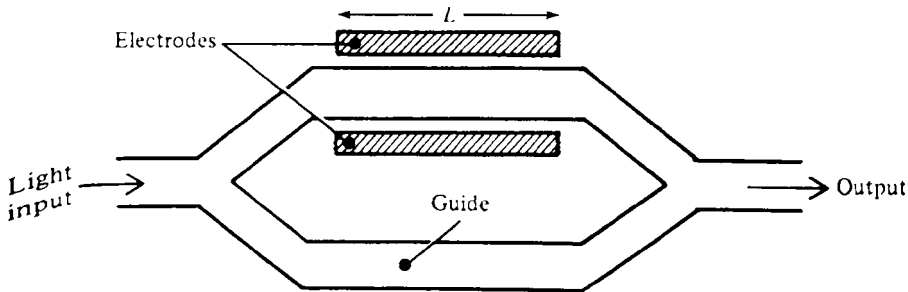


FIG. 9.38 Interferometric modulator; a voltage applied across the electrodes affects the refractive index of the upper guide over a distance L .

Some work has also been carried out into the possibility of using organic polymers in phase shift modulators. These materials offer the advantage of low cost and flexible processing techniques coupled with electro-optic coefficients which are as large, and possibly larger, than in lithium niobate. However, there are problems with high insertion losses (when used in conjunction with silica optical fibers) and the low thermal stability of the material. Semiconductor materials such as GaAs and InP do exhibit an electro-optic effect, but the electro-optic coefficients are not very large (see Table 3.1). However, semiconductor quantum well structures exhibit an electric-field-induced change in their effective refractive index which is much larger than in the bulk material.⁴ Although only some 10% of the mode field in such waveguides is actually within the quantum well structure, which reduces the effectiveness of the refractive index changes, the changes are still appreciably larger than in the bulk material. The interaction lengths required (i.e. the length L in Fig. 9.38) is thereby reduced from a few millimetres to several hundred micrometres. This has two main advantages: first it increases the device density on the substrate, thereby reducing costs; and secondly it reduces the electrical capacitance, thereby increasing the modulation bandwidth.

A different type of switch/modulator may be constructed which utilizes the coupling of energy between waveguides when they are brought into close proximity. We have met this phenomenon before when considering optical fiber couplers (section 8.5.2); it arises as a result of the overlap of the evanescent field in one guide with the core of the neighbouring guide. Using coupled mode theory it can be shown that (ref. 8.28) the amount of energy coupled into the neighbouring guide over a length z (Fig. 9.39) is proportional to the factor F_c , where, for identical waveguides,

$$F_c = \sin^2(Cz) \quad (9.27)$$

whilst for non-identical waveguides

$$F_c = \frac{C^2}{C^2 + \delta^2} \sin^2[z(C^2 + \delta^2)^{1/2}] \quad (9.27a)$$

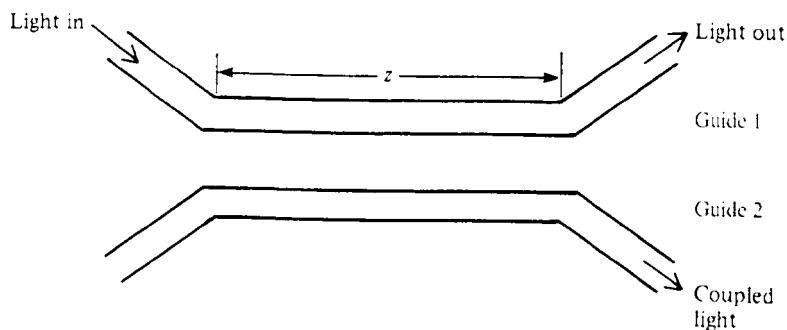


FIG. 9.39 Two waveguides in close proximity over a coupling distance z . As explained in the text evanescent field overlap can cause energy to be transferred between the guides.

Here, C characterizes the coupling between the guides and the factor δ is given by

$$\delta = \frac{\pi}{\nu C}(n_1 - n_{11})$$

where n_1 and n_{11} are the refractive indices of the two guide materials and ν is the frequency of the radiation.

If $\delta = 0$, then after a coupling length $z = L_c$ where $L_c = \pi/(2C)$, all the power in one guide will be transferred to the other, whilst after a further distance L_c all the power will have been transferred back again (Fig. 9.40a). However, if $\delta \neq 0$ then not only is less energy exchanged but also the energy change takes place more rapidly with distance (Fig. 9.40b). Suppose that we have identical waveguides which can be changed to non-identical guides in some way. If the coupling length is equal to L_c then, whilst the guides are identical ($\delta = 0$), all the energy in one guide will couple into the other. However, if the guides are now made non-identical and with δ having a value such that $L_c(C^2 + \delta^2) = \pi$ (i.e. $\delta = \sqrt{3}C$), then over a coupling length L_c no energy is exchanged. This means we can switch energy from one waveguide to another provided we can 'switch' from a situation where $\delta = 0$ to one where $\delta = \sqrt{3}C$ (see Problem 9.16).

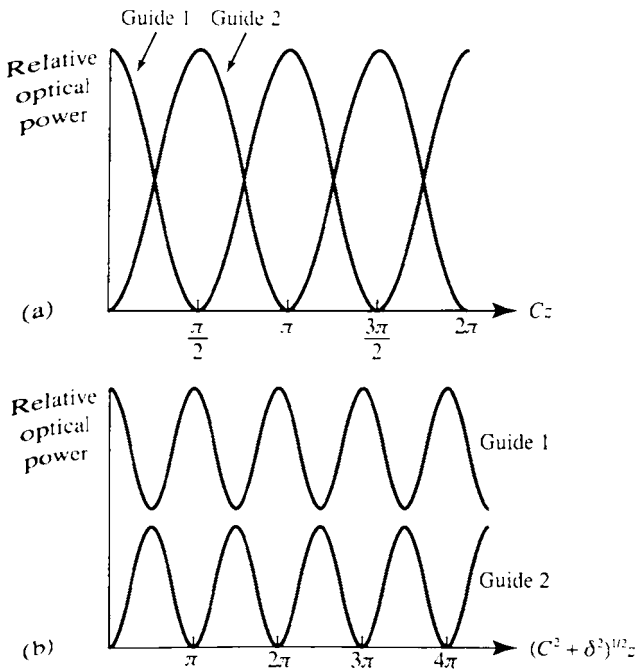


FIG. 9.40 Transfer of optical power between waveguides as a function of coupling distance z (see Fig. 9.39) according to eqs (9.27) and (9.27a). In (a) the guides are identical and all the energy is exchanged between the guides, whereas in (b) the guides are non-identical and only a partial exchange of energy is possible.

Several configurations have been built to implement this, of which the simplest is shown in Fig. 9.41. Electrodes are deposited above each waveguide and a potential is applied between them. The opposing vertical fields in the two waveguides can, if the material axes have been chosen correctly, induce opposite changes in the guide refractive indices and hence change the value of δ . The main problem with this arrangement is that it is not easy to achieve total energy transfer because of the difficulty of ensuring that the coupling length is exactly L_c . More complicated electrode configurations have been proposed to overcome these difficulties (ref. 9.18).

Devices such as filters and resonators may be realized in IO by incorporating periodic structures into optical waveguides. Consider, for example, a waveguide with a 'corrugation' etched upon its surface perpendicular to the direction of beam propagation (Fig. 9.42). This structure is encountered in the distributed feedback laser (section 6.2); it acts as a wavelength-dependent mirror, that is strong reflection occurs when $2D = m\lambda_0/n_1$, where D is the grating period, λ_0 the vacuum wavelength, n_1 the guide material refractive index and m an integer. Reflection bandwidths are usually narrow, but may be increased by 'chirping' the grating (Fig. 9.43).

Although all the devices discussed above use stripe waveguides, devices can also be based on slab waveguides. One of these is a beam deflector based on diffraction from an acoustic wave. An interdigital electrode structure deposited on a suitable acousto-optical material (see Fig. 9.44) can generate a beam of surface acoustic waves which can then serve to diffract light travelling along the guide. The angle through which the beam is diffracted may be changed by varying the frequency (and hence the wavelength) of the acoustic wave. This device is of course a version of the acousto-optic deflector discussed in Chapter 3.

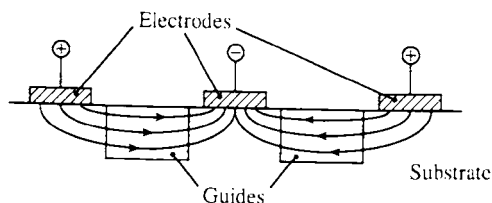


FIG. 9.41 An electrode configuration used to modify the propagation conditions in two adjacent waveguides in order to alter the coupling of radiation between them. The electric fields are in opposite directions through the guides and hence the effective refractive index of one guide will be raised whilst that of the other will be lowered.

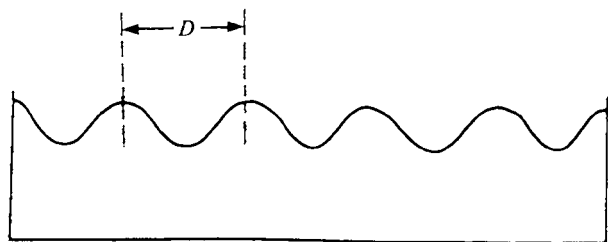


FIG. 9.42 Waveguide with a corrugation of period D etched upon it.

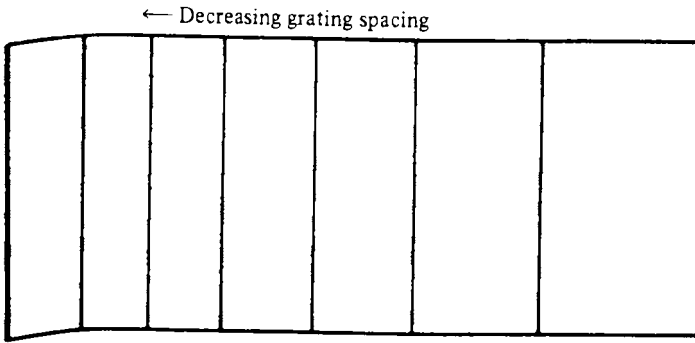


FIG. 9.43 'Chirped' diffraction grating structure. A plan view is shown with the vertical lines representing the grating peaks (or troughs).

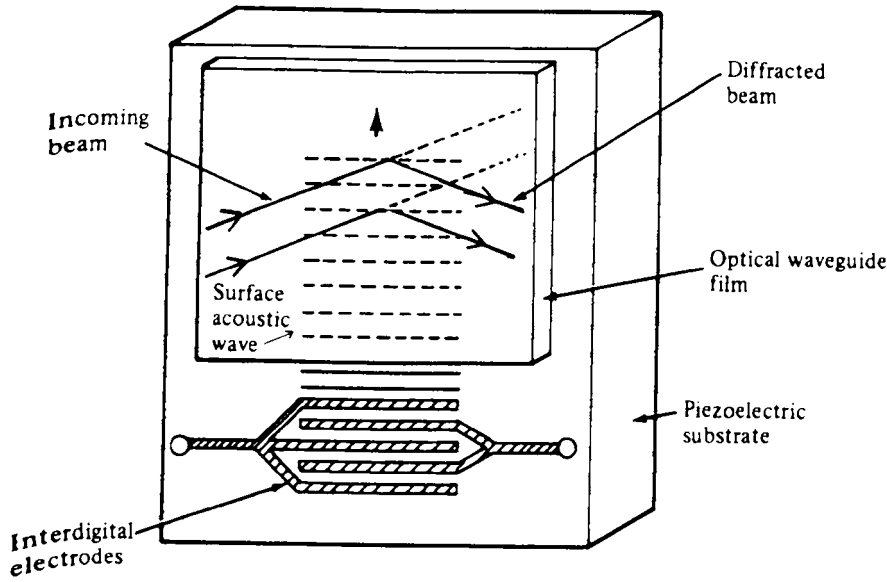


FIG. 9.44 Beam deflection using diffraction from a surface acoustic wave generated by applying an alternating voltage to an interdigital structure evaporated onto the surface of a piezoelectric substrate.

When it comes to emitters and detectors the most obvious choice for a substrate would seem to be a semiconductor. We have seen in Chapters 6 and 7 how efficient emitters and detectors can be made from them. However, many of the modulators/switches discussed above were based, not surprisingly, on materials such as lithium niobate, which exhibit relatively high electro-optic coefficients. Semiconductor materials such as GaAs and InP are electro-optically active but have appreciably smaller coefficients, as has already been pointed out. Thus the optimum substrate materials for modulators and emitters/detectors do not necessarily coincide. Since it is impossible for lithium-niobate-type materials to be made into emitters or detectors, it would appear that, if complete integration is required,

semiconductor substrates will have to be used, probably in conjunction with quantum well structure modulators which offer an increased electric field sensitivity.

In semiconductor lasers the radiation is generated within a channel of similar dimensions to those of stripe waveguides, which aids the coupling of power between them. Because of the difficulties of obtaining cleaved end mirrors, however, a distributed feedback structure (section 6.2) is usually employed, rather than the more common Fabry–Perot structure. One problem with the laser as a source is that, when not being pumped, the lasing region is absorbing; consequently arrangements are usually made for coupling the radiation from the active layer into a non-lossy guiding layer situated beneath it (see Fig. 9.45). Quantum well structures are attractive for lasers since they offer a number of advantages such as low threshold current, low temperature sensitivity and excellent dynamic behaviour as well as integrating well with modulators which are also based on quantum wells.

Finally in this section we deal with a class of devices known as ‘bistable optical devices’. In these devices two distinct optical states can exist with the possibility of switching between them. Such basic devices can form the basis of a whole series of logic gates (e.g. AND, NAND, OR and NOR gates). Furthermore if the elements can be switched by light itself then we have the basis of an all-optical computer (ref. 9.19). Such a switching device can be made by growing a multiple quantum well in place of the intrinsic region in a p–i–n photodiode (Fig. 9.46a). We suppose that light of irradiance I_{in} falls onto one face of the device and that an irradiance I_{out} emerges from the other side. The device is electrically biased using a battery and a load resistor R (and hence is known as an R-SEED). Because of the presence of the bias field the absorbance of the MQW structure is relatively small and thus for small values of I_{in} we will have that $I_{\text{out}} \approx I_{\text{in}}$. If I_{in} is now increased then so will the current flowing in the external bias circuit. As this current becomes larger so the bias actually applied to the device will fall and hence cause the absorption in the MQW region to increase (section 3.9). A point will be reached when the device suddenly switches to a state where the absorption in the MQW region is very high; a relatively large current flows in the bias circuit and the voltage across the cell is very small (which is what causes the absorption to be high). If now I_{in} is slowly reduced the high absorption state will persist for some time before the device equally suddenly switches back to the low absorption state as illustrated in Fig. 9.46(b); thus the switching process exhibits hysteresis. Suppose we

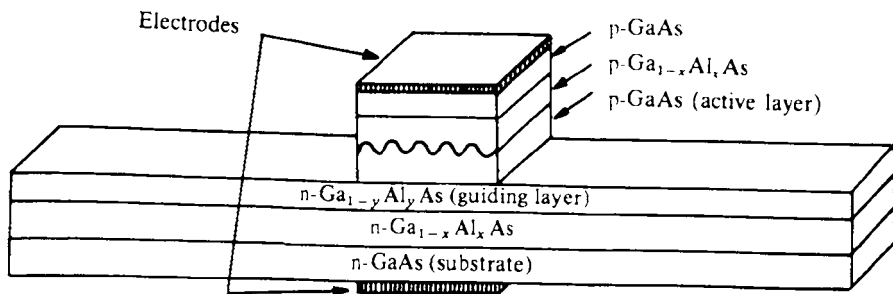


FIG. 9.45 IO semiconductor laser based on GaAs/GaAlAs using Bragg reflectors instead of cleaved end mirrors. Light from the active layer is coupled into the layer beneath, which then acts as a waveguide.

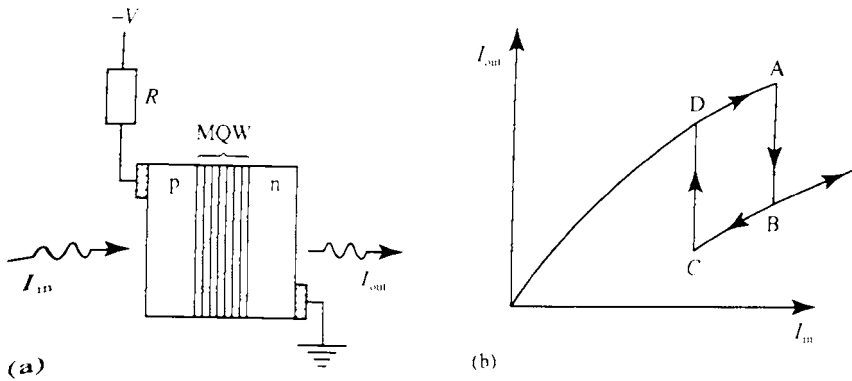


FIG. 9.46 (a) The basic structure of an R-SEED switching device. (b) The relationship between the incident (I_{in}) and transmitted (I_{out}) irradiances. As explained in the text the device exhibits a hysteresis loop.

have a bias beam of irradiance I_b where I_b is such as to be just less than the irradiance at which the device switches from low to high transmission states and that, in addition, we have a signal beam I_s where the combined irradiance $I_b + I_s$ is such as to cause the device to switch just to the low transmission state. If the device is then permanently illuminated with the bias beam, the presence or absence of the signal beam will cause the device to switch between low and high transmission states. Unfortunately it is not usually practicable to maintain the constancy of the bias beam with the required accuracy. This problem can be solved by replacing the bias resistor with another SEED device, thus generating the symmetric SEED or S-SEED (Fig. 9.47). The most important feature of the S-SEED is that changes of state occur if the ratio of the two input power irradiances alters; changes in the

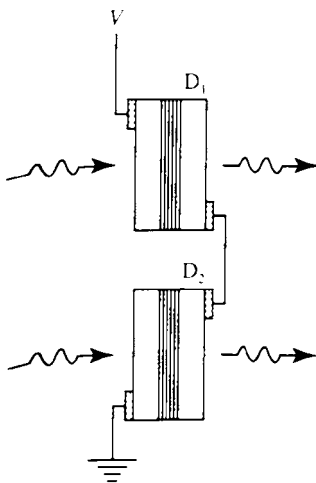


FIG. 9.47 Two SEED devices (D_1 and D_2) connected together to form an S-SEED.

absolute irradiances of the input beams do not cause a change of state. The output of the device consists of a pair of light beams, one of which will be in the 'on' or high irradiance state, the other in the 'off' or low irradiance state. In *single rail logic* the signal beam is only incident on one of the devices, but other pulses must be applied to both elements. Suppose, for example, that a logical AND operation is required, the sequence of operation is as follows:

1. Initial 'preset' pulses are applied so that device 1 is set at high ($D_1 = 1$) and device 2 at low ($D_2 = 0$).
2. The input signal pulses, S_a and S_b , are applied to device 1.
3. 'Clock' pulses are applied to both devices; the result of a logical AND operation may then be read from the output of device 2 and of a logical NAND from the output of device 1.

Suppose, for example, that $S_a = 0$ and $S_b = 1$; after S_a there will be no change in the state of the detectors, but after S_b the detectors' states will switch, so that $D_1 = 0$ and $D_2 = 1$. The clock pulse will then read a logical 1 from device 2 as required (and a logical 0 will be read from device 1, giving a NAND operation). The logical operations OR and NOR can similarly be obtained by initially setting $D_1 = 0$ and $D_2 = 1$. By combining S-SEED devices more complex devices can be constructed, for example a 2×2 switching element can be constructed by using six S-SEED elements (ref. 9.20).

9.4.3 IO devices

One of the earliest IO devices to be fabricated was an optical spectrum analyzer, designed to display the frequency spectrum of a radio-frequency (RF) signal (ref. 9.21). The layout is illustrated in Fig. 9.48. Light from a semiconductor laser is launched into a waveguide and the beam subsequently rendered parallel by the use of a 'geodesic' lens. This is a circular indentation made in the substrate layer with the guide layer thickness being unchanged across the indentation (Fig. 9.49). Such a structure behaves like a 'one-dimensional' lens (ref. 9.22). The parallel beam then passes through the acousto-optic beam deflector. If a single RF frequency is present, the amount of deflection will depend on the instantaneous value of the frequency. A second lens subsequently focuses the light onto a particular photodetector in a photodetector array. Each detector element corresponds to a particular frequency (or, more accurately, a narrow range of frequencies). If, in fact, more than one frequency is present, the light will be divided into different components that are then focused onto different elements, thus enabling the frequency spectrum of the RF signal to be obtained. The device as illustrated here is obviously a 'hybrid' since the emitters and detector elements are separate devices which are merely attached to the edge of the substrate.

Some more recent integrated devices have involved the integration of electronic components (usually gain devices based on transistors) with optical elements. For example, a Pockels-effect-type modulator can be driven from the input of a transistor amplifier so that the actual electrical input signal to the device can be relatively small. Other candidates are the drive circuits for laser emitters and the front end amplifier for an optical detector. As an example of the latter, Fig. 9.50 shows a (simplified) schematic diagram of an MSM photodiode (section

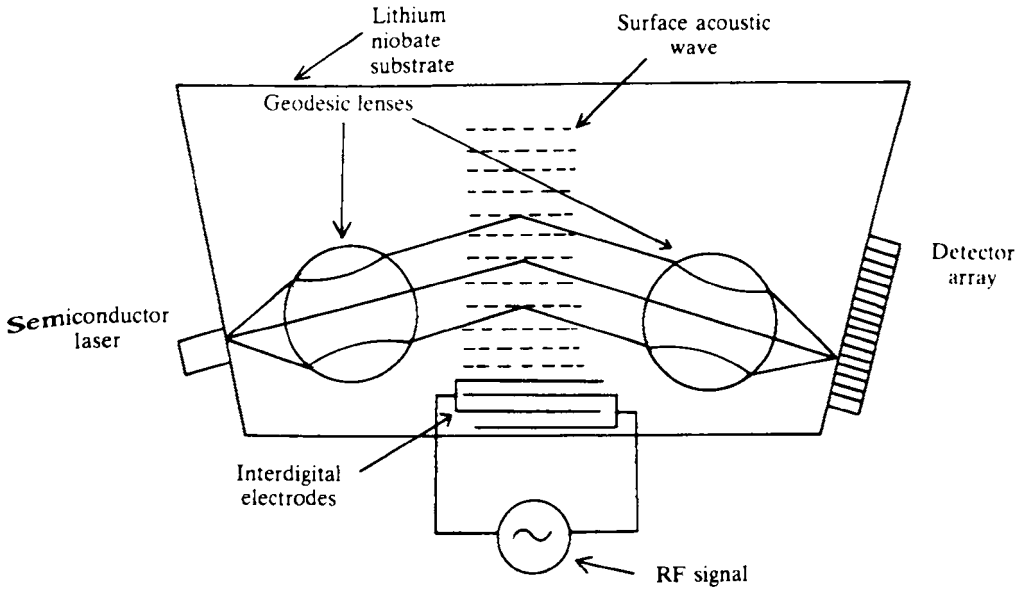


FIG. 9.48 Integrated optical spectrum analyzer based on an acousto-optic deflector.



FIG. 9.49 Cross-section of a geodesic lens in a waveguide.

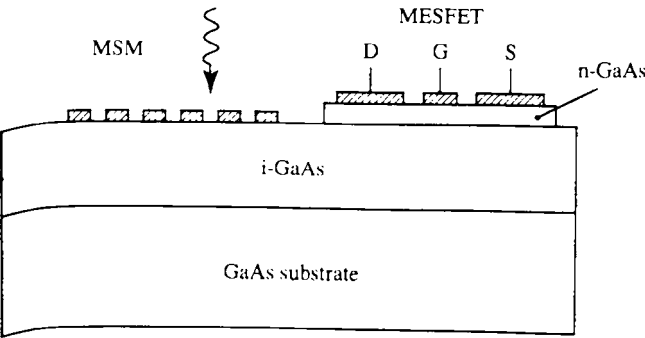


FIG. 9.50 The combination of an optical detector (MSM) and amplifier (MESFET) grown together on a single GaAs substrate.

7.3.6.5) integrated with a MESFET (*Metal–Semiconductor Field Effect Transistor*) device grown on a GaAs substrate. The GaAs MESFET is somewhat similar to the silicon-based JFET; it differs from the latter in that the gate is not a p–n junction but a Schottky barrier metal contact. When an external potential is connected between the source and drain contacts a current will flow between them through the n-type GaAs layer. If a negative voltage is applied to the gate electrode the reverse bias creates a depletion region in the n-type channel under the gate which then modulates the gate–source current. Such devices are capable of operating with modulation frequencies of up to 20 GHz. Unfortunately the reliance on Schottky barriers makes such devices impractical for use with narrow bandgap semiconductor materials such as InGaAs/InP, since then leakage currents become excessive.

Figure 9.51 illustrates an integrated laser emitter–external modulator combination for use at 1.55 μm which has been demonstrated by AT&T Laboratories. The laser, modulator and connecting waveguide are based on quantum well structures and the laser has an external Bragg reflector for feedback. Within the laser itself the eight quantum wells are based on compressively strained layers of InGaAs/InGaAsP, whereas within the modulator there are 10 layers based on InGaAsP/InP. The whole is based on an InP substrate. Other examples of communication-orientated IO devices that have been successfully demonstrated are a balanced heterodyne receiver and a tunable transmitter for wavelength division multiplexing.

Recently integrated optical devices based on the materials Si, $\text{Si}_{1-x}\text{Ge}_x$ and SiO_2 have become commercially available (ref. 9.23). Silicon is reasonably transparent over the important communication wavelength range 1.3 μm to 1.6 μm and waveguides can be made by having silicon ($n = 3.5$) as the core material with SiO_2 ($n = 1.5$) as a substrate material. The large refractive index difference between core and substrate means that single mode planar waveguides will require very small silicon core regions, but by adopting the ridge waveguide structure as shown in Fig. 9.52 mode field dimensions can be made similar to those

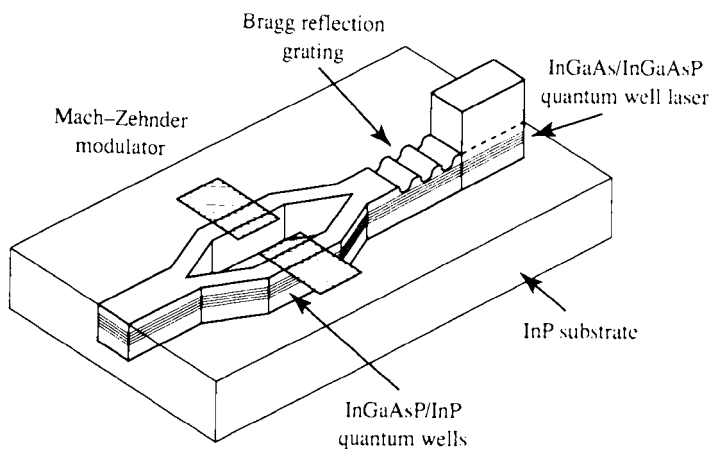


FIG. 9.51 An integrated optical emitter–modulator combination based on a quantum well semiconductor laser and a Mach–Zehnder-type modulator.

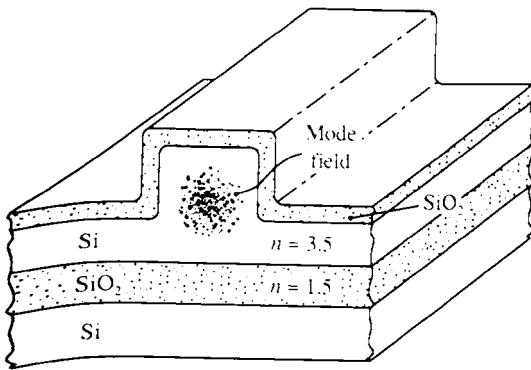


FIG. 9.52 Structure of a ridge waveguide using silicon as the guide layer.

in single mode silica fibers. As a material silicon has the advantage of a mature processing technology, for example the ability to etch V-grooves in silicon (Fig. 8.36) can be put to good use in the alignment of fibers with the waveguides. As has been pointed out previously (section 4.6.1.1) it is not possible to make efficient emitters based on silicon, although various types of modulator are possible, for example Bragg reflector types (ref. 9.24). Detectors which make use of interband transitions in $\text{Si}_{1-x}\text{Ge}_x$ are also possible, although the fact that the material is an indirect bandgap semiconductor means that relatively long lengths of detector material are needed to absorb appreciable amounts of radiation. Available commercial devices often use hybrid emitters and detectors butt coupled to the waveguides.

By no means are all applications of IO communication based: for example, there are IO versions of the active matrix LCD display (Fig. 4.31) and the read head of the compact disc player. However, the number and complexity of possible IO devices is much too broad to be adequately dealt with here and the interested reader is directed to ref. 9.25 for further information.

NOTES

1. The absorption spectrum is shifted somewhat with respect to the fluorescence spectrum and pumping at 1480 nm is possible.
2. A soliton was in fact first observed in 1832 in the form of a water wave on a canal.
3. In fact there is a loss of radiation, although to simplify the discussion we may ignore it, see ref. 9.15a.
4. The refractive index changes result from changes in the absorption coefficient when a field is applied, see section 3.9.

PROBLEMS

- 9.1 How accurately may the amplitude of a signal waveform be specified using an 8 bit number?

- 9.2 The analog modulation bandwidth required for television is about 4 MHz; what is the minimum bit rate required for a digital transmission?
- 9.3 Estimate the maximum range for a line-of-sight optical communication system using mirror collimators of 0.1 m diameter and operated at $0.85\text{ }\mu\text{m}$. The minimum power required by the detector is 10^{-9} W and the emitter launches 1 mW of radiation. Take into account losses arising from both (a) beam diffraction and (b) atmospheric attenuation (which you may assume to be 10 dB km^{-1}).
- 9.4 An LED source emits 1 mW from an area $100\text{ }\mu\text{m}$ in diameter. It is butt joined to a step index fiber with a core diameter of $50\text{ }\mu\text{m}$ and core refractive index of 1.48. The cladding has a refractive index of 1.46. Assuming the LED source is a Lambertian emitter, estimate the energy coupled into the fiber.
- 9.5 The much narrower beam divergence of lasers enables them to couple substantially more power into a fiber than LEDs. A convenient expression for their surface luminance as a function of angle is given by $B(\theta) = B(0)\cos^n\theta$. A Lambertian source is characterized by $n = 1$, whereas lasers may have $n \approx 20$ or more. Using this expression, calculate the coupling efficiency into a fiber as a function of n assuming $\text{NA} \ll 1$.
- 9.6 By assuming that the field profile in single mode fibers is Gaussian (i.e. $\mathcal{E}_0(r) = \mathcal{E}_0(0)\exp[-(r/\omega_0)^2]$), show that if a Gaussian beam is focused onto the end of the fiber so that its field profile is $\mathcal{E}_B(r) = \mathcal{E}_B(0)\exp[-(r/\omega_B)^2]$ then, ignoring any Fresnel reflection losses, the coupling efficiency, η_c , may be written

$$\eta_c = \frac{4\omega_0^2\omega_B^2}{(\omega_0^2 + \omega_B^2)^2}$$

(Hint: write eq. 9.9 in terms of radial coordinates before evaluating the overlap integral.)

- 9.7 By assuming that the mode field within a single mode fiber is Gaussian (so that $\mathcal{E}(r) = \mathcal{E}(0)\exp[-(r/\omega_0)^2]$) and by ignoring any Fresnel reflection effects, show that the coupling efficiency between two such fibers as a function of a separation z can be written as

$$\eta_c = \frac{4(1 + x^2)}{(2 + x^2)^2}$$

where $x = z\lambda_0/(\pi\omega_0)$. (Hint: use eq. 9.10 in conjunction with eq. 5.34.)

Sketch a graph of the variation of η_c with x .

In a demountable join involving single mode fiber, the fibers have a separation between their end faces of $10\text{ }\mu\text{m}$. If the fibers involved have core diameters of $8\text{ }\mu\text{m}$, numerical apertures of 0.10 and the wavelength of radiation involved is $1.3\text{ }\mu\text{m}$, estimate the resulting coupling loss.

- 9.8 An emitter can couple some 10^{-3} W of optical radiation at 900 nm into a fiber which has an attenuation of 5 dB km^{-1} . A receiver is used which requires an average of 500 photons per bit for an acceptable bit-error rate. Plot a graph of the maximum fiber length possible for bit rates of between 10^5 and 10^9 bps . Modify this diagram to include the

effects of fiber dispersion (5 ns km^{-1}) if the maximum allowable pulse spreading must not exceed $1/(2 \times \text{bit rate})$.

- 9.9** A photodiode is used in the photoconductive mode in conjunction with a load resistor of 50Ω . Optical power of $6 \mu\text{W}$ at a wavelength of 800 nm falls onto the detector. The light is amplitude modulated at frequencies up to 100 MHz . Assuming the detector has a quantum efficiency of 0.7 , and a reverse bias saturation current of 2 nA , calculate the resulting signal-to-noise ratio.

If the photodiode has a capacitance of 5 pF , determine whether the choice of a 50Ω load resistor is optimum from a signal-to-noise point of view. If it is not determine the optimum value and the corresponding signal-to-noise value.

- 9.10** Show that the optimum gain of an avalanche photodiode may be obtained by solving the equation

$$M^3 \frac{dF(M)}{dM} = \frac{2 \times \text{Johnson noise}}{\text{shot noise}}$$

where $F(M)$ is the excess noise factor. Hence, by substituting for $F(M)$ from eq. (7.38a) and for the Johnson and shot noise terms from eq. (9.11) and eq. (9.12) respectively, show that the optimum signal-to-noise ratio may be obtained by solving

$$M_{\text{opt}} \left(\frac{M_{\text{opt}}^2}{r} - \frac{1}{r} + 1 \right) = \frac{4kTF_n}{(i_A + i_D)eR_L}$$

- 9.11** Determine the optimum gain for an avalanche photodiode made from GaAs where the electron-to-hole ionization ratio, r , is 5 . Assume the ratio of amplifier to shot noise is 1000 . Estimate the improvement in signal-to-noise ratio obtained in this instance by using an APD rather than a p-i-n photodiode.

Repeat the calculation assuming an avalanche photodiode made from germanium where $r = 1$.

- 9.12** Assuming that the noise in a detector is governed by statistical fluctuations in the photon arrival rate, calculate the minimum signal power level at a wavelength of $1.3 \mu\text{m}$ required to achieve a BER of 10^{-12} when the bit rate is 5 Gbps .

- 9.13** Show that the magnitude of the output signal voltage of the transimpedance amplifier of Fig. 9.21 can be written

$$V_{\text{out}} = \frac{i_A R_f}{(1 + 4\pi^2 f^2 C_j^2 R_f^2 / G^2)^{1/2}}$$

where G is the gain of the operational amplifier. Assume that the value of G is large and remains independent of frequency.

- 9.14** It is required to construct an optical fiber communication link 5 km long with an analog modulation bandwidth of 10 MHz . The proposed system uses multimode step index fiber with core and cladding refractive indices of 1.48 and 1.46 respectively. At the operating wavelength the fiber exhibits an attenuation of 3 dB km^{-1} . The emitter is an

LED which can launch 200 mW of optical power into the fiber and the detector a p-i-n photodiode which requires -40 dBm to achieve the required signal-to-noise ratio. No splices are required. The detector and emitter have rise times of 100 ns and 10 ns respectively. Determine whether the link should operate satisfactorily and if not discuss the component changes you would recommend.

- 9.15 A waveguide modulator of the Mach-Zehnder type for digital modulation is to be made based on lithium niobate as a substrate material for use with radiation where $\lambda_0 = 1.3 \mu\text{m}$. The guides themselves are $2 \mu\text{m}$ wide and the voltage to be applied is 5 V. Estimate the length of waveguide over which the voltage must be applied. Assume that the transverse Pockels effect coefficient in lithium niobate has the value $30 \times 10^{-12} \text{ m V}^{-1}$, and that the ordinary refractive index is 2.2.
- 9.16 A modulator is to be made based on the coupling of radiation between two closely spaced identical waveguides. With no voltage applied the coupling length for $\lambda_0 = 0.8 \mu\text{m}$ is 10 mm, and this is the actual length of the coupling region in the device. In this condition radiation launched into one guide will be wholly coupled into the other guide at the output. Voltages are applied across the waveguides over this length so that the transverse fields produced within the guides act oppositely to each other. If the waveguides are $4 \mu\text{m}$ wide, determine the approximate magnitude of the voltage which will cause the radiation to switch back to the original guide at the output. Assume that the transverse Pockels effect coefficient in lithium niobate has the value $30 \times 10^{-12} \text{ m V}^{-1}$, and that the ordinary refractive index is 2.2.

REFERENCES

- 9.1 T. G. Hodgkinson, D. W. Smith, R. Wyatt and D. J. Mayon, 'Coherent optical fiber transmission systems', *Br. Telecommun. Technol. J.*, **3**, 5, 1985.
- 9.2 H. Taub and D. L. Schilling, *Principles of Communication Systems*, McGraw-Hill, New York, 1971, Chapter 5.
- 9.3 J. M. Senior, *Optical Fiber Communications* (2nd edn), Prentice Hall, Hemel Hempstead, 1992, Section 11.6.3.
- 9.4 *Ibid.*, Chapter 13.
- 9.5 H. F. Wolf (ed.), *Handbook of Fibre Optics*, Granada, St Albans, 1979, Chapters 8 and 11.
- 9.6 R. M. Gagliardi and S. Karp, *Optical Communications*, Wiley-Interscience, New York, 1976, Chapter 2.
- 9.7 G. Einarsson, *Principles of Lightwave Communications*, John Wiley, Chichester, 1996, Section 5.5.
- 9.8 J. Gowar, *Optical Communication Systems* (2nd edn), Prentice Hall, Hemel Hempstead, 1993, Chapter 24.
- 9.9 S. D. Personick, *Fiber Optics: Technology and Applications*, Plenum Press, New York, 1985, Chapter 9.
- 9.10 (a) S. Matsushita, K. Kawai and H. Uchida, 'Fibre-optic devices for LAN applications', *J. Lightwave Technol.*, **3**, 554, 1985.
- (b) P. E. Green, *Fiber Optic Networks*, Prentice Hall, Englewood Cliffs, NJ, 1993.

- 9.11 H. Ishio, J. Minowa and N. Kiyoshi, 'Review and status of wavelength division multiplexing technology and its applications', *J. Lightwave Technol.*, **2**, 448, 1984.
- 9.12 D. M. Sprit and M. J. O'Mahony, *High Capacity Optical Transmission Explained*, John Wiley, Chichester, 1995.
- 9.13 M. Nakazawa, E. Yamada and K. Suzuki, '10Gb/s soliton data transmission over one million kilometers', *Electron. Lett.*, **27**, 1270-2, 1991.
- 9.14 D. B. Anderson, *Optical and Electro-optical Information Processing*, MIT Press, Cambridge, MA, 1965, p. 221.
- 9.15 (a) R. G. Hunsperger, *Integrated Optics: Theory and Technology* (3rd edn), Springer-Verlag, Berlin, 1991.
(b) T. Tamir (ed.), *Integrated Optics* (2nd edn), Springer-Verlag, Berlin, 1979.
- 9.16 (a) D. L. Lee, *Electromagnetic Principles of Integrated Optics*, John Wiley, New York, 1986.
(b) R. G. Hunsperger, *op cit.*, Chapter 4.
- 9.17 H. Nishihara, M. Haruna and T. Suhara, *Optical Integrated Circuits*, McGraw-Hill, New York, 1985, Section 2.5.1.
- 9.18 T. Tamir (ed.), *Guided-Wave Optoelectronics* (2nd edn), Springer-Verlag, Berlin, 1990, pp. 179-80.
- 9.19 (a) A. D. McAulay, *Optical Computer Architectures*, John Wiley, New York, 1991.
(b) J. Jahan, *Optical Computing Hardware*, Academic Press, San Diego, 1994.
- 9.20 H. E. Escobar, 'All-optical switching systems near practical use', *Laser Focus World*, Oct., 135-41, 1994.
- 9.21 T. J. Joseph, T. R. Ranganath, J. Y. Lee and M. Pedicott, 'Performance of the integrated spectrum analyser', *Proc. SPIE*, **321**, 134, 1982.
- 9.22 B. Chen and O. S. Ramer, 'Diffraction limited lens for integrated optical circuits', *IEEE J. Quantum Electron.*, **15**, 853, 1979.
- 9.23 B. Schuppert *et al.*, 'Integrated optics in silicon and silicon germanium heterostructures', *J. Lightwave Technol.*, **14**, 2311-23, 1996.
- 9.24 C.-C. Wang *et al.*, 'Ultrafast all-silicon light modulator', *Opt. Lett.*, **19**, 1453-5, 1994.
- 9.25 (a) J. T. Boyd (ed.), *Integrated Optical Devices and Applications*, IEEE Press, New York, 1991.
(b) R. Waynant and M. Ediger (eds), *Electro-Optic Handbook*, McGraw-Hill, New York, 1994, Chapters 26 and 27.

Non-communications applications of fibers

Although the main driving force in the development of optical fibers has been their use for telecommunications purposes, other, perhaps more mundane, uses promise to be of equal, if not greater, importance. In this chapter we examine some of these.

10.1 Optical fiber sensors

In many industrial processes there is a need to monitor such quantities as displacement, pressure, temperature, flow rate, liquid level and chemical composition. Ideally, the measurement technique should be reliable, robust, corrosion resistant, intrinsically safe and free from external interference. Obviously, optical fibers and indeed optical methods in general have the potential for making a significant contribution in this area. In many cases, however, well-developed alternative technologies already exist and it may be some time before optical fibers can make any really significant penetration into the market.

Optical fiber sensors themselves can be divided into two main categories, namely 'intrinsic' and 'extrinsic'. In the latter, the modulation resulting from a change in the measurand takes place outside the optical fiber, which acts merely as a convenient transmission channel for the radiation. In intrinsic sensors, on the other hand, the quantity to be measured acts directly on the fiber itself to modify the radiation passing down the fiber.

As far as intrinsic sensors are concerned, it is possible to modulate the amplitude, the phase or the state of polarization of the radiation in the fiber. In multimode fibers, however, mode coupling and the usually random relationships between the phases and polarization states of the propagating modes generally preclude the use of either phase or polarization modulation. Thus, intrinsic multimode fiber sensors almost invariably involve amplitude modulation. On the other hand, with single mode fibers both phase and polarization modulation become possible. Phase modulation opens up the possibility of fiber-based interferometric sensors that can offer exceptionally high sensitivities. It is convenient therefore to treat multimode and single mode intrinsic fiber sensors separately.

10.1.1 Multimode extrinsic optical fiber sensors

Some of the simplest extrinsic fiber-based sensors are concerned with the measurement of movement or position. For example, when two fiber ends are moved out of alignment, the coupling loss depends on the displacement. The actual relationships between loss and dis-

placement were considered in section 8.5 (see also Problem 8.17). A similar type of sensor uses a shutter moving between two fiber ends that are laterally displaced (Fig. 10.1a). Improvements in sensitivity are possible by placing a pair of gratings within the gap, one fixed, the other movable (Fig. 10.1b). Here, however, although the sensitivity has increased, the range has decreased, since the output will be periodic in the spacing of the grating. The range of movement possible for the single shutter sensor is obviously limited by the fiber core diameter. If a beam expander is employed between the fibers (Fig. 10.1c), then the range can be greatly expanded. The displacement--transmission characteristic of the expanded beam system is considered in Problem 10.1.

One of the first commercially available displacement sensors was the 'Fotonic sensor'. This uses a bundle of fibers, half of which are connected to a source of radiation, the other half to a detector (Fig. 10.2a). If the bundle is placed in close proximity to a reflecting surface then light will be reflected back from the illuminating fibers into the detecting fibers. The amount detected will depend on the distance from the fiber ends to the surface. To analyze this dependence, we consider the somewhat simpler situation where there are

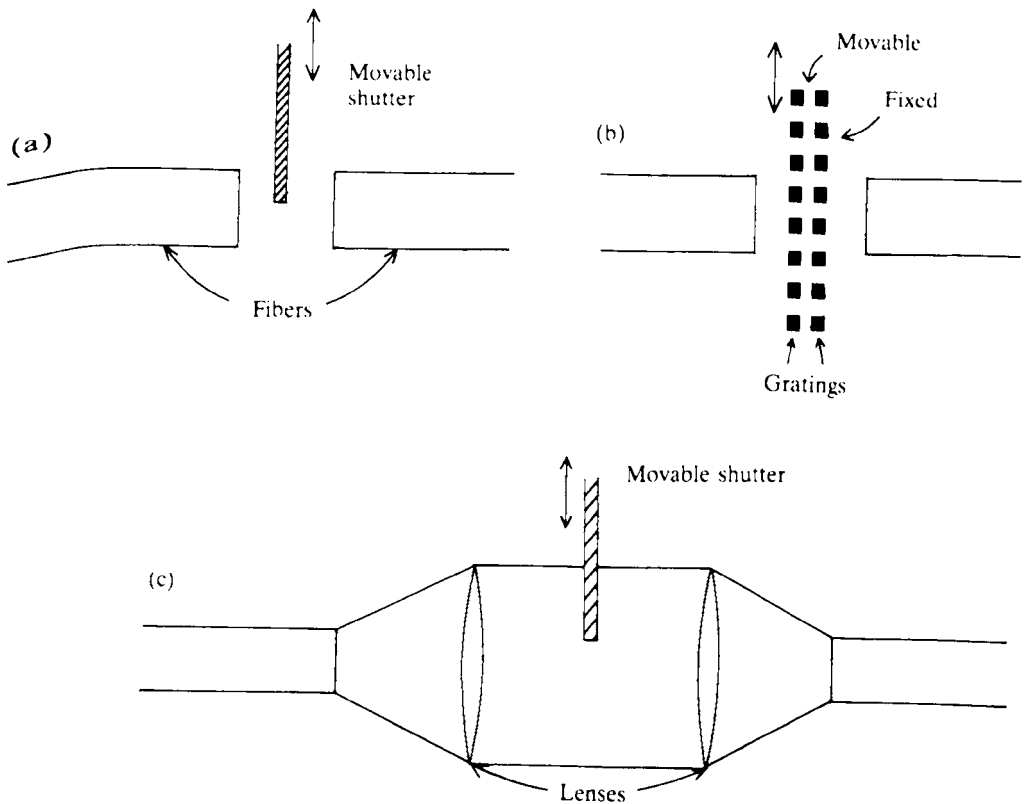


FIG. 10.1 Simple displacement sensors. In (a) a movable shutter varies the light coupled between two longitudinally displaced fibers. In (b), the use of two gratings increases the sensitivity. In (c), a beam expansion system enables an increase in the range of measurable displacements to be increased.

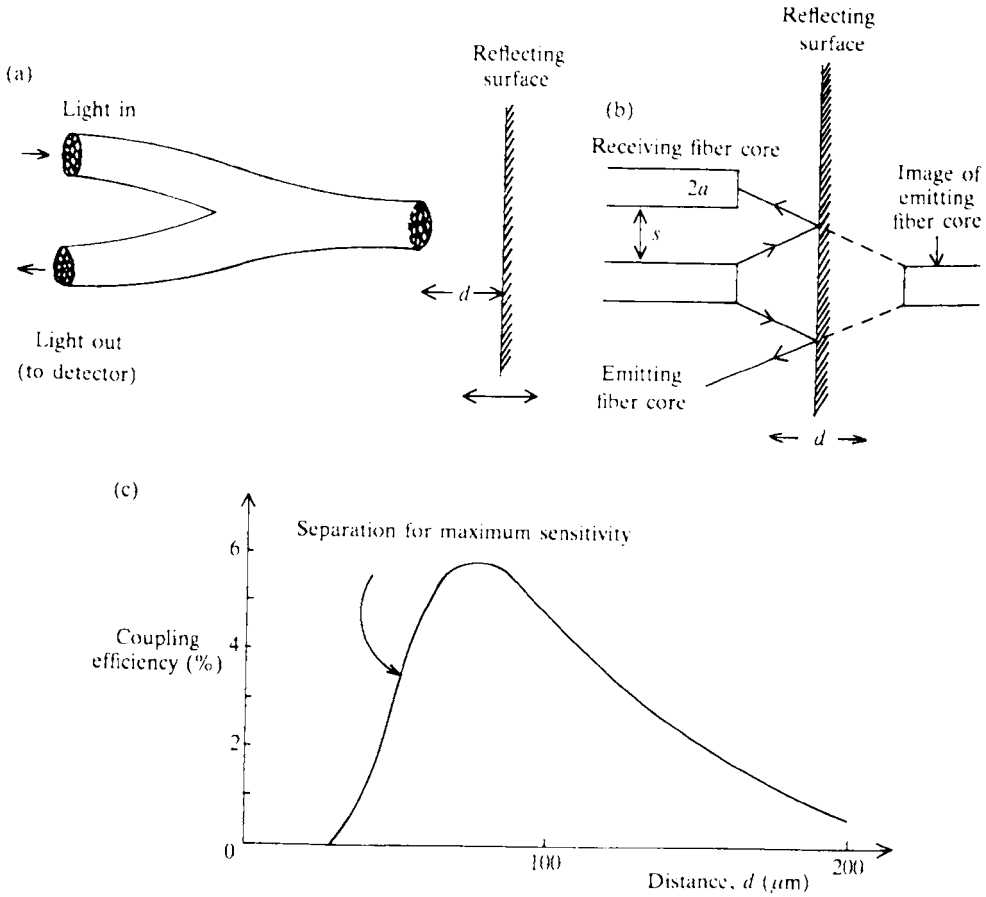


FIG. 10.2 Illustration of the Fotonic sensor. The general layout using fiber bundles is shown in (a). A two-fiber version is shown in (b), which can be used to derive the form of the output-distance (d) relationship. The typical result of such a calculation taking $a = 100 \mu\text{m}$, $s = 100 \mu\text{m}$ and $\text{NA} = 0.4$ is shown in (c).

just two fibers. If we regard the reflecting surface as a mirror, the problem then reduces to that of the coupling between two fibers that are displaced both laterally and longitudinally (Fig. 10.2b). The form of the relationship between displacement and light output may be determined by considering the overlap between the sensing fiber core area and the cross-section of the light cone emitted by the image of the emitting fiber. We can readily appreciate that at very small fiber-surface distances, no light will be coupled between the two fibers. Then, beyond a certain critical distance, there will be an increasing overlap between the above areas and the coupled radiation will increase rapidly. Once the detecting fiber area is completely filled, however, the output will fall with increasing distance. At large distances, an inverse square law will then be obtained. The displacement output characteristic is considered in more detail in Problem 10.2, and a typical result of this analysis is shown in Fig. 10.2(c).

In practice, when a fiber bundle is used instead of just two fibers the displacement–output characteristic will be somewhat different and will depend on how the emitting and receiving fibers are distributed (usually randomly), but the overall shape remains similar to that of Fig. 10.2(c). Because of the very non-linear nature of the curve, the sensor is not very suitable for the measurement of large displacements, although it is possible to increase the range by using a lens system. In fact, the sensor was developed originally for non-contact vibration analysis.

Any displacement measurement technique is readily adapted to the measurement of pressure, and those mentioned above are no exception. For example, the Fotonic sensor may be placed close to a reflecting diaphragm with a constant pressure maintained on the sensor side. Any change in external pressure will cause flexing of the diaphragm and a consequent change in the instrument's output. It should be remembered, however, that none of the instruments described above is linear except over a very limited range of displacements. Accurate calibration over the whole range is therefore required.

As well as displacement/pressure sensors, a number of extrinsic fiber temperature sensors have been proposed. For example, the bandgap of semiconductors such as GaAs is temperature dependent (Fig. 10.3a) and a simple sensor can be made in which a piece of the semiconductor is placed in the gap between the ends of two fibers (Fig. 10.3b). Light with a wavelength corresponding approximately to the semiconductor bandgap is sent down one of the fibers and the power emerging from the other is measured and can be related to the temperature.

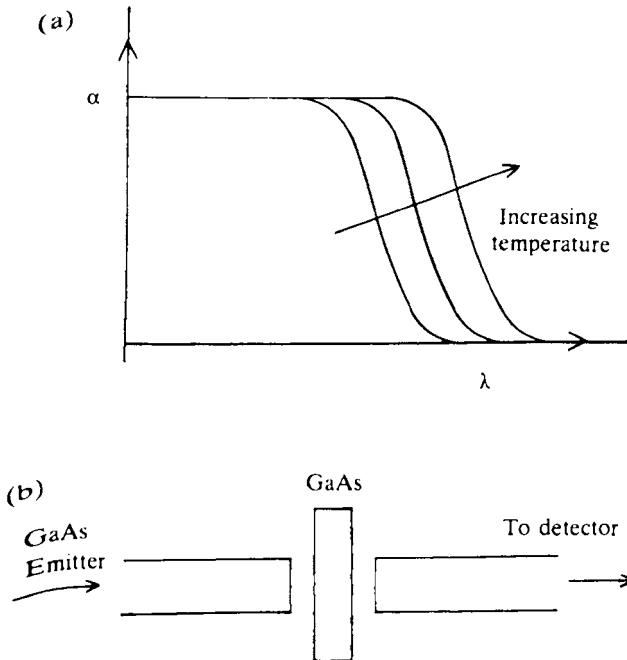


FIG. 10.3 (a) Schematic variation of the absorption coefficient (α) of GaAs with both wavelength (λ) and temperature. (b) Temperature sensor utilizing variations in the transmission of GaAs with temperature.

Another temperature sensor, the Fluoroptic sensor, is available commercially and is claimed to have a sensitivity of 0.1°C over the range -50°C to 250°C . The instrument relies on the temperature variation of the fluorescence in europium-doped lanthanum oxysulfide ($\text{Eu}:\text{La}_2\text{O}_3\text{S}$). A small amount of this material is placed on the object whose temperature is to be measured, and the fluorescence is excited by illuminating it with ultraviolet light transmitted down a fairly large diameter ($400\text{ }\mu\text{m}$) plastic-coated silica fiber. The source of radiation is a quartz-halogen lamp whose output has been filtered to remove any unwanted higher wavelengths. Another, similar, fiber picks up some of the emitted fluorescence and carries it back to the detector system (Fig. 10.4). In fact, the phosphor emits at more than one wavelength and it is the intensity ratio of two of the lines which is measured. Because a ratio is measured, any fluctuations in the irradiance of the source are not important.

At the detector end of the fiber, the radiation is split into two using a beam splitter. Each beam falls onto a silicon photodiode having an appropriate filter in front to isolate the particular wavelength required. The ratio of the two signals then provides the temperature information, which is usually contained in a 'look-up' table. Because the ultraviolet output from a quartz-halogen lamp is small and the fiber absorption relatively large at short wavelengths, the output of the phosphor is quite small. Efficient detectors and low noise preamplifiers are required, and the maximum fiber length is restricted to about 15 m. Nevertheless, the

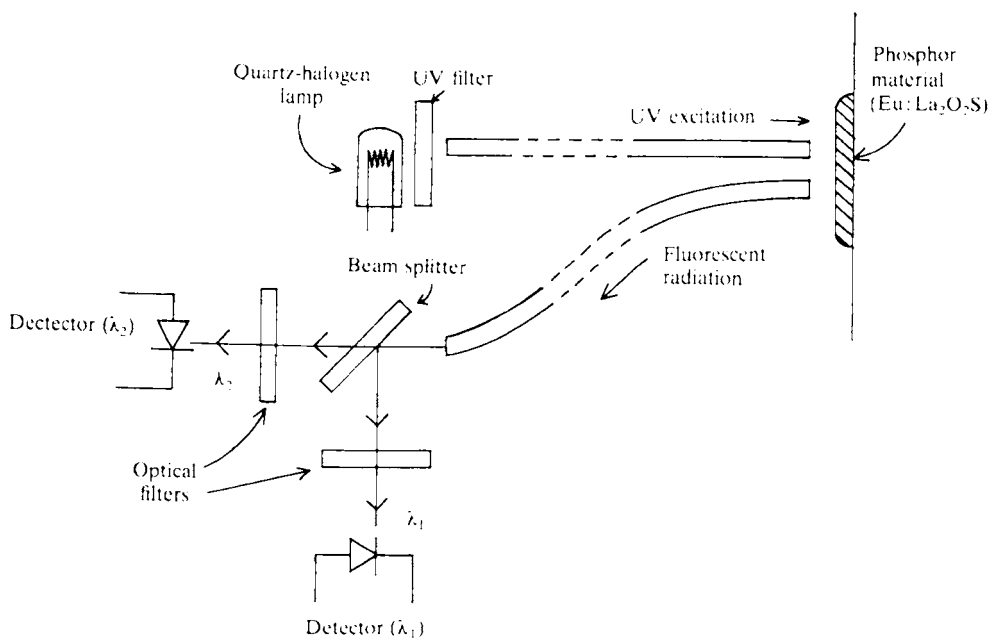


FIG. 10.4 Schematic layout of the Fluoroptic temperature sensor. The fluorescent radiation generated in the phosphor is separated into its two main constituent wavelengths (λ_1 and λ_2) and the relative optical power of these wavelengths is determined by using a beam splitter followed by two optical bandpass filters to isolate the two wavelengths.

device provides a performance superior to thermocouples, and allows point temperature measurements in semiremote hostile environments.

10.1.2 Multimode intrinsic optical fiber sensors

One of the ways in which we can influence the amount of radiation flowing down a fiber is by means of microbending loss (section 8.4.1), and this can therefore be made the basis of a displacement or pressure transducer. In a typical device, the fiber passes between a pair of ridged plates which impart a periodic perturbation to the fiber. In fact, we have met such an arrangement before in the guise of a mode scrambler (section 8.6). If step index fiber is used, a particular periodic perturbation of wavelength Λ will only couple together a few modes (see eq. 8.30). However, it may be shown (ref. 10.1) that, with a graded index fiber where the profile parameter, α , is equal to 2, all modes are coupled together when

$$\Lambda_c = \frac{2\pi a}{\sqrt{2\Delta}} \quad (10.1)$$

Typically $\Lambda_c \sim 1$ mm (see Example 10.1), and with this value for Λ we expect very high microbending loss and hence high sensitivity for the sensor.

EXAMPLE 10.1 Microbending sensor periodicity

We consider the periodicity required for maximum microbending loss in a graded index fiber. We assume the fiber has a $50\text{ }\mu\text{m}$ core diameter and maximum core and cladding refractive indices of 1.48 and 1.46 respectively. We have therefore that $a = 25\text{ }\mu\text{m}$ and that (from Example 8.5) $\Delta = 0.0135$. Hence from eq. (10.1)

$$\begin{aligned} \Lambda &= \frac{2\pi \times 25 \times 10^{-6}}{(2 \times 0.0135)^{1/2}} \text{ m} \\ &= 0.96 \text{ mm} \end{aligned}$$

When the modes in a fiber are excited by a coherent source, they are capable of interfering with each other and thus of producing an interference pattern across the end of the fiber. The pattern obtained will depend on the phase differences developed between the modes as they travel along the fiber, which is impossible to predict. Provided there are no perturbations acting on the system, however, the pattern should remain unchanged. If now the fiber is slightly flexed in any way, mode coupling will change the distribution of energy amongst the modes, and hence produce a change in the interference pattern across the fiber end. Of course, unless there is a significant amount of coupling into lossy modes, there will not be any great change in the total amount of energy emerging from the fiber. If, however, we consider only a small portion of the whole area of the fiber end, any change in the interference pattern as a whole is almost certain to produce quite significant changes in the emerging energy. Thus, if a detector is so placed as to intercept only a small portion of the total light emerging from the fiber, its output should vary when there is any deformation of the fiber.

By its very nature, such a detector will be very non-linear, though in some circumstances this may not be a great disadvantage. For example, by laying the fiber just below ground level it may be possible to detect the presence of intruders, since their footsteps will cause deformation of the fiber. All that is required is for the output to trigger an alarm when the change in the signal exceeds some predetermined level.

The Bragg fiber grating structure (section 8.7.5.2) can be used as a very useful sensing element. It will be recalled that the grating will reflect radiation of wavelength λ_B which satisfies the equation

$$\lambda_B = 2mn_1\Lambda \quad (10.2)$$

where $m = 1, 2, 3$, etc., Λ is the periodicity of the grating and n_1 the refractive index of the core. The exact value of the product $n_1\Lambda$ will depend on both temperature and strain within the fiber. As far as temperature changes are concerned both the grating wavelength and the refractive index will be affected by temperature and we can write

$$\Delta\lambda_B = 2m\Lambda n_1 \left(\frac{1}{n_1} \frac{dn_1}{dT} + \frac{1}{\Lambda} \frac{d\Lambda}{dT} \right) \Delta T$$

This can be written as

$$\Delta\lambda_B = \lambda_B(\beta + \alpha)\Delta T \quad (10.3)$$

where $\beta = (dn_1/dT)/n_1$ and α is the linear expansion coefficient. Similarly the application of strain (ϵ) will affect both the grating spacing and the refractive index (via the photoelastic effect), and we may write (ref 10.2)

$$\Delta\lambda_B = \lambda_B(1 - p_e)\Delta\epsilon \quad (10.4)$$

where p_e is an effective photoelastic coefficient given by

$$p_e = \frac{n_1^2}{2} [(1 - \mu)P_{12} - \mu P_{11}]$$

where P_{11} and P_{12} are Pockels coefficients and μ is Poisson's ratio. Example 10.2 calculates the resulting sensitivities for silica fiber.

EXAMPLE 10.2 Bragg fiber grating sensitivity

For pure silica the thermal expansion coefficient (α) is $5 \times 10^{-7} \text{ K}^{-1}$ whilst the quantity $(dn_1/dT)/n_1$ (i.e. β) is $6.8 \times 10^{-6} \text{ K}^{-1}$. Thus according to eq. (10.3) the wavelength sensitivity to temperature changes of the fiber grating structure, assuming $\lambda_B = 1.55 \text{ }\mu\text{m}$, is

$$\Delta\lambda_B = 1.55 \times 10^{-6} (5 \times 10^{-7} + 6.8 \times 10^{-6}) = 0.0113 \text{ nm K}^{-1}$$

As far as sensitivity to strain is concerned we have that $P_{11} = 0.126$, $P_{12} = 0.274$ and $\mu = 0.17$. It then follows that

$$p_e = \frac{1.46^2}{2} [(1 - 0.17) \times 0.274 - 0.17 \times 0.126] = 0.229$$

Thus as far as strain is concerned

$$\Delta\lambda_B = 1.55 \times 10^{-6}(1 - 0.229) = 1.2 \times 10^{-6} \text{ m } \epsilon^{-1} \quad \text{or} \quad 1.2 \times 10^{-3} \text{ nm } \mu\epsilon^{-1}$$

There are a number of ways in which the Bragg sensor can be 'interrogated' to obtain a measure of the reflection wavelength λ_B . For example, if radiation from a tunable laser is incident on the grating and its output wavelength scanned across the appropriate wavelength range then strong reflection will be obtained at λ_B . The magnitude of the back-reflected radiation is easily monitored using the set-up illustrated in Fig. 10.5. Several such sensors may be employed at different positions along the fiber provided that the wavelength ranges associated with the sensors are mutually exclusive and also that the laser scanning range is sufficiently large. The wavelength of each sensor is determined by correlating the time at which a reflected pulse is detected to the laser wavelength at that time. In another measurement technique radiation from a broad-band source is sent down the fiber. Light at wavelength λ_B will be removed from the beam, leading to a 'notch' appearing in the transmitted spectrum and either the reflected or transmitted spectrum can be analyzed to obtain λ_B . However, the changes in λ_B are small and difficult to measure directly except with costly instruments such as the optical spectrum analyzer. A number of possible measurement schemes have been proposed most of which involve matching the wavelength λ_B to the resonant wavelength of some other optical system such as a Fabry-Perot interferometer.

Bragg grating sensors offer a number of advantages over other types. For example, they offer a relatively high resolution of strain or temperature, the output is a linear function of the measurand and they are insensitive to fluctuations in light intensity. In addition they are relatively easy to fabricate and do not compromise the structural integrity of the fiber.

We know from our discussion of fibers in Chapter 8 that the mode field extends out into the cladding. If, therefore, over a certain portion of a fiber we have a cladding material which changes its optical properties (e.g. refractive index or absorption coefficient) in response to

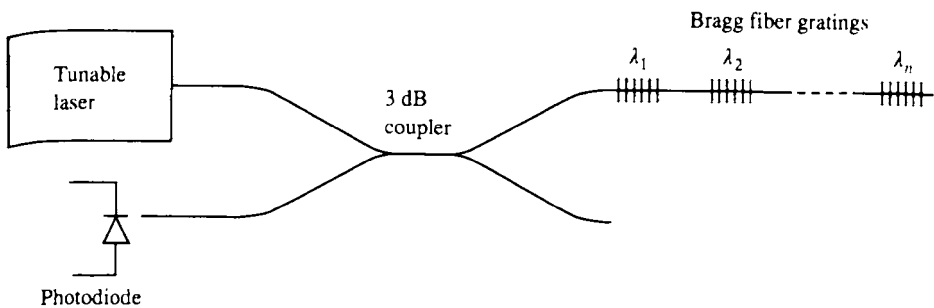


FIG. 10.5 Illustration of a technique that can be used to interrogate an array of Bragg fiber diffraction gratings. The photodiode will only receive a signal when the output of the laser corresponds to one of the reflection wavelengths ($\lambda_1, \lambda_2, \dots, \lambda_n$).

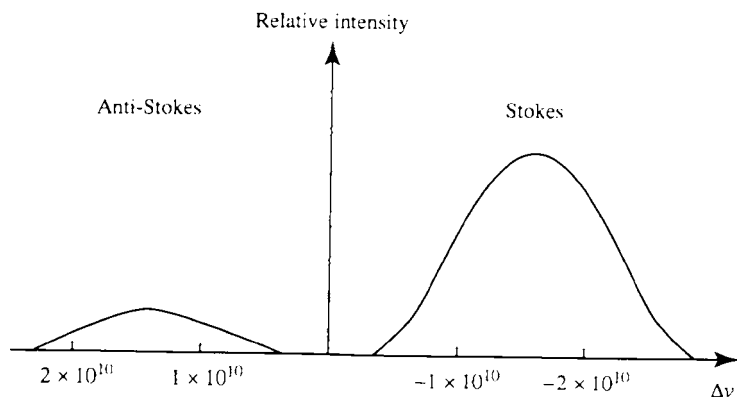


FIG. 10.6 Raman scattering spectrum in silica; the scattered frequency differs from the incident frequency by an amount $\Delta\nu$.

changes in the surroundings, then the mode field will be affected to some extent. Sensors relying on this basic principle have been made to measure liquid refractive indices and various ionic concentrations and pH values (ref. 10.3).

10.1.3 Distributed fiber sensors

Most of the sensors discussed so far have been concerned with the measurement of physical parameters at a particular point in space. However, there are a number of situations where it is highly desirable to monitor a physical parameter continuously along a length; whilst it is always possible to use a large number of point sensors this makes for a very cumbersome (and costly) solution. A much better idea is to use a single fiber and to monitor some property dependent on the required measurand at the required number of points along the fiber. The flexibility of fiber makes it relatively easy to install over a chosen measurement path and thus allows retrospective fitting. The value of having access to the spatial/temporal behaviour of strain and temperature in, for example, large critical structures such as dams, aircraft, spacecraft, bridges, etc., is clear both from the point of view of safety monitoring and for a better understanding of behaviour under unexpected conditions.

The principle of the OTDR (section 8.6.6) provides a means of achieving the spatial resolution; thus the time delay between a launch pulse and a returning pulse is readily converted into the distance along the fiber to where the returning pulse was generated. In the OTDR the radiation received originates in Rayleigh scattering, which in silica is relatively insensitive to temperature. Some early demonstrations used liquid core fibers where the scattering is much more temperature dependent. However, such fibers are not generally suitable for practical applications. Another possibility is the use of fibers that have been doped with rare earth elements (e.g. the Er^{3+} -doped fibers used in optical amplifiers), in which some of the absorption features vary with temperature and hence can influence the gradient of the

OTDR trace. For example, a holmium-doped fiber has shown a temperature sensitivity of about 1°C with a spatial resolution of about 4 m.

Other sensors have been developed based on scattering mechanisms which involve the interaction of the radiation with the vibrations of the lattice. In terms of quantized vibrations these may be described as photon-phonon interactions. There are two main types of lattice vibration (or phonon) involved: one in which neighbouring atoms are moving with (almost) the same phase, giving rise to *acoustic* phonons; and the other in which neighbouring atoms are moving almost 180° out of phase, giving rise to *optical* phonons. Optical phonons can only exist in materials where the unit cell contains more than one atom. During the interaction process the photon can either create or absorb a phonon, so that the energy of the photon is changed. If a phonon is created the photon energy will decrease and hence also its frequency; in this case we talk of a *Stokes* shift. When a phonon is absorbed, on the other hand, the photon energy and frequency will both increase and we have an *anti-Stokes* shift. The scattering arising from optical phonons is called *Raman* scattering, whilst that from acoustic phonons is called *Brillouin* scattering. Since optical phonons have energies which are several orders of magnitude larger than acoustic phonons, the frequency shifts involved are also correspondingly larger. Another difference is that the magnitude of the shift in the case of Brillouin scattering depends on the scattering angle of the photon; the frequency shift attains its maximum value for backward scattering and falls to zero for forward scattering. Because the material we are mainly concerned with (silica) is amorphous, the phonons have quite a wide range of frequencies, resulting in a relatively wide frequency spectrum of the scattered radiation, particularly so for Raman scattering (Fig. 10.6) where a Stokes frequency shift of about 1.5×10^{13} Hz is obtained with a linewidth of a similar magnitude to the shift itself. In the case of Brillouin scattering a maximum frequency shift of about 10^{10} Hz with a linewidth of about 3×10^7 Hz is observed.

In Raman scattering the ratio of the Stokes to anti-Stokes scattering irradiancies, R_s , depends on temperature according to (ref. 10.4)

$$R_s = \left(\frac{\lambda_s}{\lambda_a} \right)^4 \exp \left(-\frac{h\Delta\nu}{kT} \right) \quad (10.5)$$

where λ_s and λ_a are the wavelengths at which Stokes and anti-Stokes scattering takes place and $\Delta\nu$ is the difference in frequency between the exciting radiation and anti-Stokes line. R_s changes by about 0.7% per $^\circ\text{C}$ over the range 1 to 100°C (see Example 10.3). Thus a measurement of R_s can be used as the basis of a (distributed) temperature sensor. Because a ratio is involved, the measurement is not affected by changes in irradiance of the probe beam. One problem with the technique is that Raman scattering is, at the wavelengths involved here, about a factor 10^{-3} weaker than Rayleigh scattering; consequently multi-mode fibers are normally used in conjunction with high power semiconductor lasers. Temperature resolutions of $1-2^\circ\text{C}$ with spatial resolutions of 1–2 m over distances of 5–10 km have been obtained (ref. 10.5). Both forward and backscattering are possible, with forward scattering being more easy to implement since the launch and scattered pulses emerge from opposite ends of the fiber, and hence pulse launch and detection is much easier to implement although pulse timing information has to be coordinated between the two ends of the fiber.

EXAMPLE 10.3 Temperature sensitivity of Raman scattering sensor

From eq. (10.5) we have

$$\frac{dR_s}{dT} = \left(\frac{\lambda_s}{\lambda_a}\right)^4 \exp\left(-\frac{h\Delta\nu}{kT}\right) \left(-\frac{h\Delta\nu}{kT^2}\right)$$

so that

$$\frac{\Delta R_s}{R_s} \approx -\frac{h\Delta\nu}{kT^2} \Delta T$$

Taking the Raman shift in silica to be about 1.5×10^{13} Hz and assuming that $T = (273 + 50)$ K = 323 K, we have that the fractional change in R_s for a 1°C change is

$$\frac{6.6 \times 10^{-34} \times 1.5 \times 10^{-13}}{1.38 \times 10^{-23} \times (323)^2} = 6.88 \times 10^{-3} \text{ per } ^\circ\text{C} \quad \text{or} \quad 0.7\% \text{ per } ^\circ\text{C}$$

Brillouin scattering may also be used, although the technique is rather more complicated than in the case of the Raman sensor. In the usual configuration light from a CW laser is launched into one end of a single mode fiber, whilst a train of laser pulses is launched into the other end. The emission frequency of the pulsed laser is higher than that of the CW laser by the Brillouin shift. The pulsed laser now acts as a pump beam for *stimulated Brillouin emission*, so that the CW beam will increase in irradiance as it passes through the fiber. Now in fact the Brillouin frequency shift depends on temperature, so that if the temperature of the fiber is non-uniform then gain will only take place in those parts of the fiber which are at a particular temperature (which can then be related to the difference between the laser emission frequencies). If the power of the CW beam is monitored following the launch of a pump pulse, an increase will be observed whenever Brillouin gain occurs. Positional information about where the gain is taking place can be obtained from the time delay between the pump pulse launch and the arrival of the increased gain in the CW beam. If the frequency difference between the lasers is altered then different temperature regions will be picked out. Thus by scanning the frequency of one of the lasers the temperature profile along the fiber can be mapped out. Rather longer lengths of fiber can be employed (i.e. several tens of kilometres) than in the case of Raman scattering sensors with similar temperature but slightly lower spatial (≈ 5 m) resolutions (ref. 10.6). Unlike Raman scattering, Brillouin scattering is sensitive to strain as well as temperature and hence simultaneous measurements of both these quantities are possible.

When it becomes possible to monitor rapidly and accurately the state of a structure it is then possible to envisage some sort of feedback mechanism whereby the form or the structure is modified to suit the prevailing conditions; thus an aircraft wing may change shape to suit varying heights, speeds and turn radii. This is the idea of *smart structures* and the advantages of optical fiber such as low mass, short response time and freedom from electromagnetic interference make it a very serious candidate for this type of situation (ref. 10.7).

10.1.4 Single mode fiber sensors

In single mode fiber sensors, we are mainly dealing with the effects of the external quantities to be measured on the phase (or mode velocity) of the light within the fiber. Some sensors involve polarization modulation, but these will be discussed later.

10.1.4.1 Phase-modulated sensors

One obvious way in which we may detect phase changes is to construct an interferometric system where the phase of the beam through a sensing fiber is compared with that of a reference beam. The Mach-Zehnder configuration is easily realized using optical fibers (Fig. 10.7). Light from a semiconductor laser is launched into a single mode optical fiber and the radiation split and sent into 'sensing' and 'reference' fibers. The radiation from these two fibers is then mixed and split again at another 3 dB coupler; the output from each fiber is then detected.

We now consider the form of the output expected from the interferometer. The electric fields of the two beams at one of the detectors may be written as $A_s \exp[i(\omega t + \Delta\phi)]$ and $A_r \exp(i\omega t)$ where $\Delta\phi$ represents the phase difference between the two. In general, $\Delta\phi$ will be made up of a 'static' differential phase term $\Delta\phi_d$ and a signal term ϕ_s where $\Delta\phi = \Delta\phi_d + \phi_s$. The total field is the sum of these two fields and the detector output (R_1) is proportional to the product of this sum and its own complex conjugate, that is

$$R_1 \propto \{A_s \exp[i(\omega t + \Delta\phi)] + A_r \exp(i\omega t)\} \{A_s \exp[-i(\omega t + \Delta\phi)] + A_r \exp(-i\omega t)\}$$

or

$$R_1 \propto A_r^2 + A_s^2 + A_r A_s [\exp(i\Delta\phi) + \exp(-i\Delta\phi)]$$

$$R_1 \propto A_r^2 + A_s^2 + 2A_r A_s \cos \Delta\phi$$

Assuming for simplicity that $A_r = A_s = A$, then

$$R_1 \propto 2A^2 [1 + \cos(\Delta\phi)] \quad (10.6)$$

This response as a function of $\Delta\phi$ is shown in Fig. 10.8, which illustrates the basic problem with any interferometric sensor, namely that its output is periodic in the phase difference. It is also evident that the greatest sensitivity to small variations in signal is obtained when

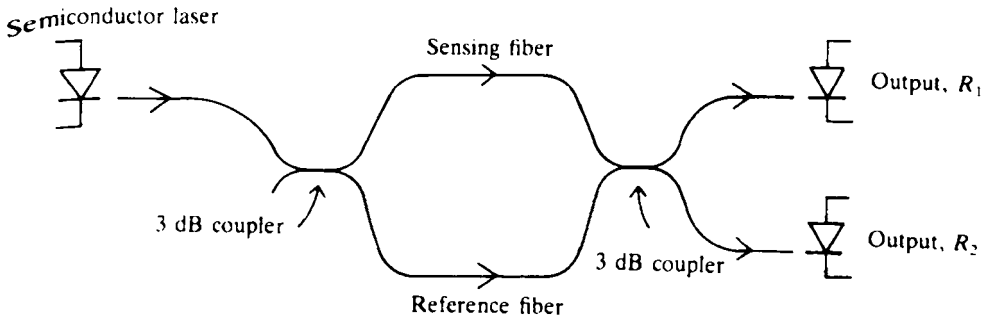


FIG. 10.7 Configuration of the optical fiber Mach-Zehnder interferometer.

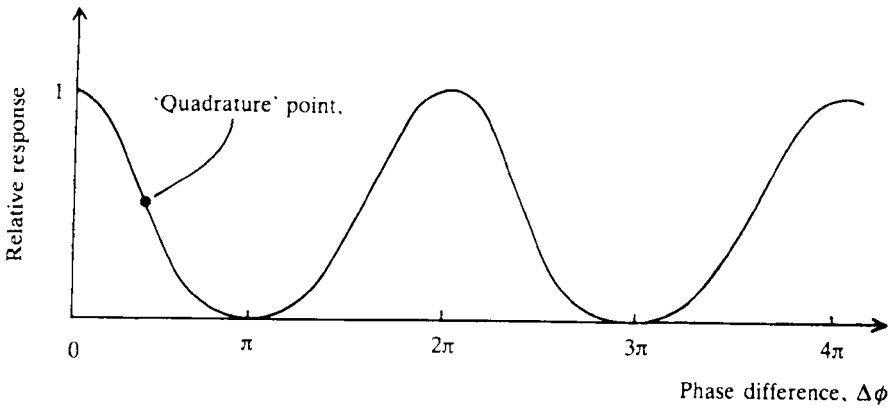


FIG. 10.8 Relative response of a Mach-Zehnder interferometer as a function of the phase difference between the two arms ($\Delta\phi$).

the system is operating at points half way between the maximum and minimum values of R_1 . These are known as the 'quadrature' points. Midway between these maximum sensitivity points the sensitivity falls to zero. Thus, even if we could ensure that we started off operating at a quadrature point it is quite likely that changes in ambient conditions around the reference arm would cause $\Delta\phi$ to change and hence cause a drift away from the quadrature point.

So far, we have considered the output from only one of the signal detectors (R_1). It may be shown (see e.g. ref. 10.8) that there is a phase difference between the beams falling on the two detectors of π . Thus, the response of the second detector (R_2) may be written as

$$R_2 \propto 2A^2[1 + \cos(\Delta\phi + \pi)]$$

or

$$R_2 \propto 2A^2[1 - \cos(\Delta\phi)]$$

Thus

$$(R_1 - R_2) \propto 4A^2 \cos(\Delta\phi) \quad (10.7)$$

In effect, therefore, by taking the difference between the two output signals we double the sensitivity of the interferometer. There are also advantages to be gained when the amplitudes of the beams from the signal and reference arms are not equal (see Problem 10.4). Problems associated with phase drift in the reference arm still remain, and we now examine ways of dealing with them.

The most straightforward technique is known as the 'active phase tracking homodyne' scheme. This requires that we control the phase of the reference arm. One of the simplest ways of doing this is to wrap the fiber around a cylinder of a piezoelectric material. When a voltage is applied to the cylinder it expands radially, thus stretching the fiber and hence inducing a phase change. If this induced phase change is written as $\Delta\phi_m$, we may hold the system at the quadrature point by requiring that $\Delta\phi_m + \Delta\phi_d = (2m + 1)\pi/2$. In the absence of any modulation of the sensing arm, this may be achieved by making the differential signal

part of a feedback loop connected to the reference arm modulator, and then ensuring that the differential signal remains close to zero (Fig. 10.9).

In the presence of a signal, we cannot use the differential signal directly in the feedback loop in this way, or the signal information will also be cancelled out. However, it is likely that any ambient modulation of the reference arm will take place at a lower frequency than that of the signal, so that the difference signal ($R_1 - R_2$) integrated over a time long compared with the signal modulation but short compared with the ambient modulation can often provide a suitable feedback signal.

The signal obtained before the feedback electronics will now be the same as in a 'perfect' system operating at the quadrature point and with no reference arm phase drift. Such a mode of operation is known as a low gain-bandwidth product (GBP) mode. One problem with this is that the output is only linear in $\Delta\phi$ when $\Delta\phi \ll \pi$. In addition, phase excursions greater than 2π can lead to ambiguities because of the periodicity of the output. A way of avoiding these difficulties is to operate the above system in a high GBP mode. That is, we remove the integrating circuit and operate the feedback circuit at frequencies high enough to include the signal frequency. At first sight, this would seem to obliterate the signal entirely; however, it

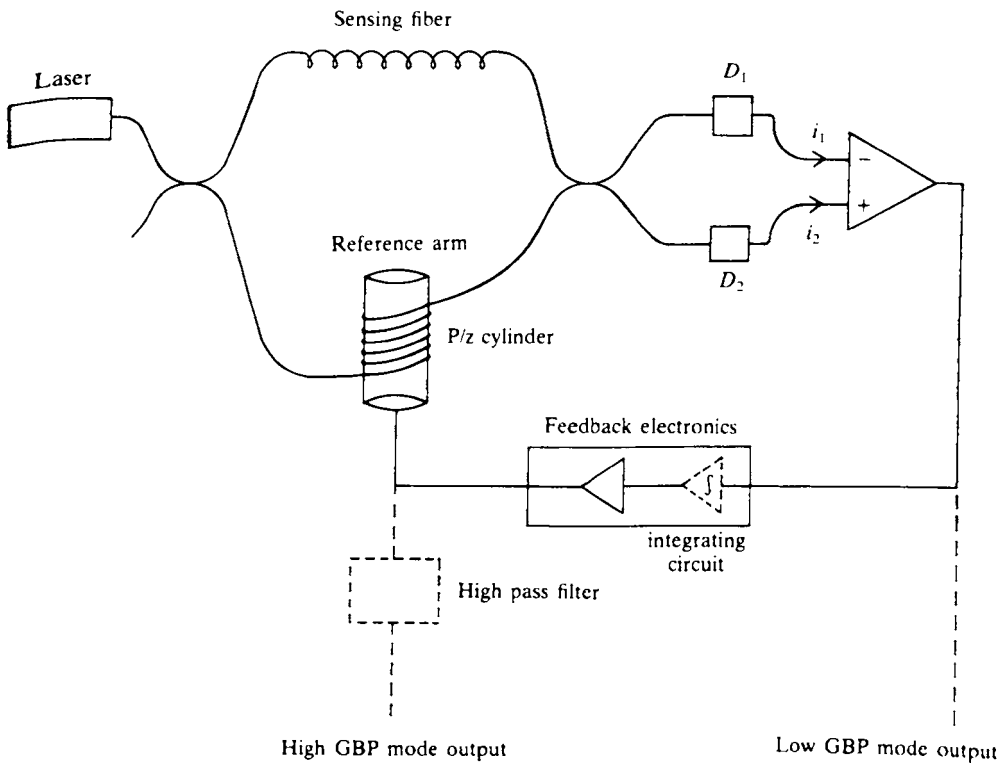


FIG. 10.9 Active phase-tracking homodyne detection using either the low gain-bandwidth product (GBP) mode or the high GBP mode. In the latter case, an integrating circuit is not present in the feedback electronics.

is still present in the signal applied to the phase modulator in the reference arm. We may now write for operation at the quadrature point

$$\Delta\phi_d + \phi_s + \Delta\phi_m = (2m + 1) \frac{\pi}{2}$$

or

$$\Delta\phi_m = (2m + 1) \frac{\pi}{2} - \Delta\phi_d - \phi_s$$

Assuming therefore that $\Delta\phi_m$ is proportional to the feedback voltage applied to the modulator, the feedback signal is linear in $(\Delta\phi_d + \phi_s)$. This linearity will extend over many 2π phase changes, the only limitation being the phase range that can be covered by the modulator itself. The required phase signal ϕ_s may be separated out using electrical filters since ϕ_s usually oscillates at a higher frequency than $\Delta\phi_d$.

The active phase tracking homodyne scheme outlined above is easy to implement and usually extremely linear in operation. There are a number of drawbacks, however. For example, the scheme uses an electrically active element which may require several hundred volts for the piezoelectric element. This may well reduce the sensor's usefulness as an electrically 'safe' device. Furthermore, although the phase range is much in excess of 2π it is still limited, and large phase excursions may necessitate resetting the system with consequent loss of information.

Many other schemes have been implemented to overcome these two basic difficulties. Nearly all involve a considerable increase in complexity and none is really ideal. (There is insufficient space to consider them here, and the interested reader may consult ref. 10.9 for more details.)

We turn now to the effects which various quantities to be measured have on the phase of the radiation in a single mode fiber. For simplicity, we assume that the mode behaves like a ray travelling down the centre of the fiber with velocity c/n_1 . Thus the phase (in radians) associated with a length L of fiber may be written as

$$\phi = \frac{2\pi L n_1}{\lambda_0}$$

If X represents the quantity to be measured, then the change in ϕ caused by a change ΔX in X may be written as

$$\Delta\phi = \frac{2\pi L}{\lambda_0} \frac{dn_1}{dX} \Delta X + \frac{2\pi n_1}{\lambda_0} \frac{dL}{dX} \Delta X \quad (10.8)$$

Because we have assumed that the internal ray angles (θ) are 90° , any changes in the fiber core radius (a) do not affect ϕ . However, even if $\theta < 90^\circ$, it can be shown (see Problem 10.5) that ϕ is much less sensitive to changes in a than it is to changes in L and n . Consequently, we may safely neglect the dependence of ϕ on a . We now estimate the sensitivity of ϕ to variations in axial force, pressure and temperature.

An axially applied force, F , will extend the fiber thus changing L (it also changes n , but

we ignore this). If Young's modulus for the fiber is E , then we have

$$\frac{\Delta L}{L} = \frac{F}{AE}$$

where A is the fiber cross-sectional area. Thus

$$\Delta\phi_F = \frac{2\pi n_1 L F}{\lambda_0 A E} \quad (10.9)$$

For silica, $E = 2 \times 10^{11}$ Pa and, taking an external fiber diameter of 100 μm , a core refractive index of 1.45 and light of wavelength 1 μm , eq. (10.9) gives

$$\begin{aligned} \Delta\phi_F &= \frac{2\pi \times 1.45}{10^{-6} \times \pi \times (10^{-4})^2 / 4 \times 2 \times 10^{11}} \\ &= 6 \times 10^3 \text{ rad N}^{-1} \text{m}^{-1} \end{aligned}$$

If the fiber is subject to a uniform radial pressure P , the fiber radius will decrease and its length will increase. We neglect the effects of any radius change, but the length change is then given by

$$\frac{\Delta L}{L} = \frac{2\nu P}{E}$$

where ν is Poisson's ratio, and hence

$$\Delta\phi_P = \frac{4\pi n_1 L \nu P}{\lambda_0 E} \quad (10.10)$$

For silica, $\nu \sim 0.2$ and hence taking the same fiber parameters as above we obtain

$$\begin{aligned} \Delta\phi_P &= \frac{4\pi \times 1.45 \times 0.2}{10^{-6} \times 2 \times 10^{11}} \\ &= 1.8 \times 10^{-5} \text{ rad Pa}^{-1} \text{m}^{-1} \end{aligned}$$

Both refractive index and fiber length are affected by temperature T , and so we may write

$$\Delta\phi_T = \frac{2\pi L}{\lambda_0} \left(\frac{dn_1}{dT} + n_1 \alpha \right) \Delta T \quad (10.11)$$

where α is the linear expansivity.

For a typical fiber

$$\frac{dn}{dT} = 7 \times 10^{-6} \quad \text{and} \quad \alpha = 5 \times 10^{-7} \text{ K}^{-1}$$

Again taking a wavelength of 1 μm and a core refractive index of 1.45 we have

$$\begin{aligned} \Delta\phi_T &= \frac{2\pi}{10^{-6}} (7 \times 10^{-6} + 1.45 \times 5 \times 10^{-7}) \\ &= 49 \text{ rad K}^{-1} \text{m}^{-1} \end{aligned}$$

In this case, the thermally induced change in the refractive index has a much larger effect than does the thermal expansion term. We may also note that in practical terms the temperature sensitivity is extremely high when compared with the effects of pressure and axial force. Assuming that the interferometer can be operated with a resolution of 10^{-6} rad, then the corresponding temperature sensitivity is 2.3×10^{-8} K. It would appear that we have the potential for a very accurate thermometer. Unfortunately, every part of the interferometer, including the reference arm, has the same high sensitivity. To maintain the whole of the reference arm at the same temperature to within 10^{-8} K is a formidable task. In fact, the high sensitivity to temperature remains a difficulty when measuring other parameters and often limits the sensitivities that can be achieved.

Quantities other than force, pressure and temperature can be measured, but usually by indirect means (rotation sensing is an exception which will be dealt with separately). For example, attaching the sensing fiber to a magnetostrictive element (whose dimensions change in the presence of a magnetic field) enables magnetic fields to be measured. One problem with this technique is that such materials do not usually respond linearly to a magnetic field. Typical sensitivities are about 10^{-7} T m $^{-1}$.

The system limitations on sensitivity are rather too complex to go into in any detail here. The ultimate limits will be determined by shot noise at the detectors (see section 7.3.3.2). Generally speaking, in fiber interferometers using low-powered lasers (i.e. 1 mW), shot noise considerations limit the minimum detectable phase shift to $10^{-6} - 10^{-7}$ rad Hz $^{-1}$. Other factors that may be important are intensity fluctuations in the laser output and laser frequency jitter. The latter will only give rise to a noise signal if the optical path lengths in the two arms of the interferometer are unequal (if they were exactly equal, of course, the phase difference would always be zero). Although it may not always be practicable to have nearly equal arm lengths, to achieve 10^{-6} rad sensitivity requires that the path differences be less than about 1 mm.

10.1.4.2 Fiber optic gyroscope

The fiber optic gyroscope relies on the Sagnac effect, that is the phase shift induced between two light beams travelling in opposite directions round a fiber coil when the coil is rotating about an axis perpendicular to the plane of the coil. To calculate the phase shift expected, we assume we have a single circular turn of fiber of radius R . It turns out that we must also assume that both light beams are effectively travelling in a vacuum. If there is no rotation, then both beams will return to their starting points in a time t , where $t = 2\pi R/c$. If, however, the ring is rotating in a clockwise direction at a rate of Ω rad s $^{-1}$, then the counterclockwise beam will arrive at its starting point sooner since the effective velocity of the beam will be $c + R\Omega$ and the time taken, t' , will be given by

$$t' = \frac{2\pi R}{c + R\Omega}$$

Similar arguments for the clockwise beam give a transit time of t'' , where

$$t'' = \frac{2\pi R}{c - R\Omega}$$

The difference between the two transit times $\Delta t_s (=t'' - t')$ is then

$$\begin{aligned}\Delta t_s &= 2\pi R \left(\frac{1}{c - R\Omega} - \frac{1}{c + R\Omega} \right) \\ &= \frac{4\pi\Omega R^2}{c^2 - R^2\Omega^2}\end{aligned}$$

Since $c^2 \gg R^2\Omega^2$,

$$\Delta t_s = \frac{4\Omega A}{c^2}$$

where $A = \pi R^2$ is the ring area.

In terms of a phase shift

$$\phi_s = 2\pi\Delta t_s\nu$$

where ν is the frequency of the light, or

$$\phi_s = \frac{8\pi\Omega A}{\lambda_0 c}$$

This expression is independent of the medium in which the light is travelling. Most optical fiber Sagnac interferometers are constructed with many turns (N say) wrapped round a circular former. We then have

$$\phi_s = \frac{8\pi\Omega AN}{\lambda_0 c} \quad (10.12)$$

EXAMPLE 10.4 Phase shift in a Sagnac gyroscope

Suppose we have 1000 turns of fiber on a former of 0.1 m radius; we may calculate the phase shift induced between two counter-rotating beams by the earth's rotation. The earth's rotation rate is about 15° per hour or $15\pi/(180 \times 3600) \text{ rad s}^{-1}$.

Assuming that $\lambda_0 = 1 \mu\text{m}$, eq. (10.12) then gives a phase shift of

$$\begin{aligned}\phi_s &= \frac{8\pi \times 15\pi \times 10^3 \times \pi(0.1)^2}{180 \times 3600 \times 10^{-6} \times 3 \times 10^8} \\ &= 1.9 \times 10^{-4} \text{ rad}\end{aligned}$$

The basic arrangement of the Sagnac fiber interferometer is shown in Fig. 10.10. The form of the output signal from the detector as a function of the rotation rate (and hence the phase change) will again be as shown in Fig. 10.8. At low rotation rates, the sensitivity will be very low and will reach maximum values at the quadrature points (i.e. when $\phi_s = (2n + 1)\pi/2$). Another problem to be surmounted involves what is known as 'reciprocity'. This concerns the necessity to ensure that the two counterpropagating beams travel absolutely identical

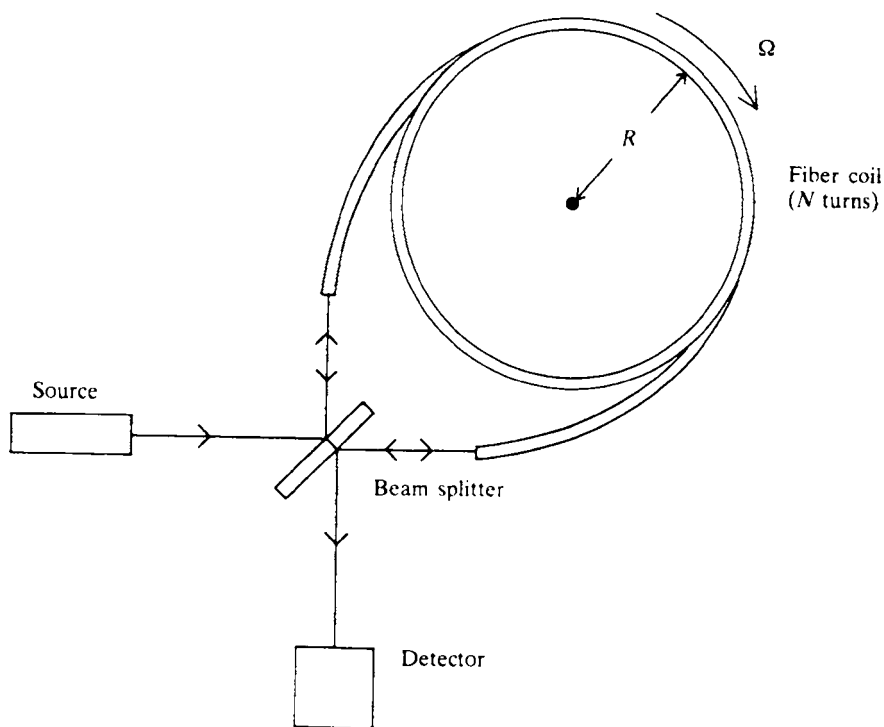


FIG. 10.10 Basic layout of the Sagnac fiber interferometer.

paths, since any path difference is a potential source of phase noise. In fact, the simple arrangement shown in Fig. 10.10 is not reciprocal; the clockwise beam is reflected twice at the beam splitter whereas the counterclockwise beam is transmitted twice through it. The reciprocity requirements can be very stringent. For example, in a system capable of measuring rotation rates as low as 10^{-4} degrees per hour the paths must be reciprocal to 1 part in 10^{16} ! Problems also arise with circularly cored fibers since they are not very good at maintaining polarization. Thus, although the two counterpropagating beams with identical polarizations may be launched into the fibers, energy may be coupled from one polarization mode into the orthogonal one as the beams traverse the fiber. The energy in the orthogonal modes will not be able to interfere with the original beams and a reduction in output will result which may be interpreted as a phase signal.

Although it might be expected that a highly coherent source would be needed in any interferometric system, this is not always the case. For example, in the present system any scattering within the fiber (e.g. Rayleigh scattering) will give rise to spurious beams, which will be able to interfere with the two main beams at the detector if the coherence length of the radiation is the order of the fiber length. With a low coherence source, on the other hand, only those beams which have nearly identical path lengths (such as the two main beams) can interfere. A 'superluminescent' diode (section 5.10.2.7) is often used as a source in fiber gyros. This is basically a semiconductor laser into which is introduced sufficient loss to prevent full laser action. It acts as a high radiance LED. However, because of its laser-like

structure it is able to couple significantly more radiation into single mode fibers than can an ordinary LED.

It can be appreciated that, with all these many problems to be overcome, the fiber gyro may never be capable of the highest levels of performance. Its place may well lie in lower cost systems of moderate resolution.

10.1.4.3 Polarimetric sensors

In polarimetric sensors, the two beams in the interferometer are provided by the two orthogonally polarized modes in a Hi-Bi fiber (section 8.7.5.1). Because of the nature of such fibers, there are two fixed directions relative to the fiber along which the polarization vectors must lie. These are referred to as the 'fast' and the 'slow' axes, depending on the relative mode velocities. If we illuminate the fiber with light polarized at 45° to these directions (Fig. 10.11), then both modes will be excited equally. An increasing phase difference will then develop as the modes travel down the fiber. The polarimetric sensor depends on the effect of external parameters on this phase difference. One problem is how to get the two

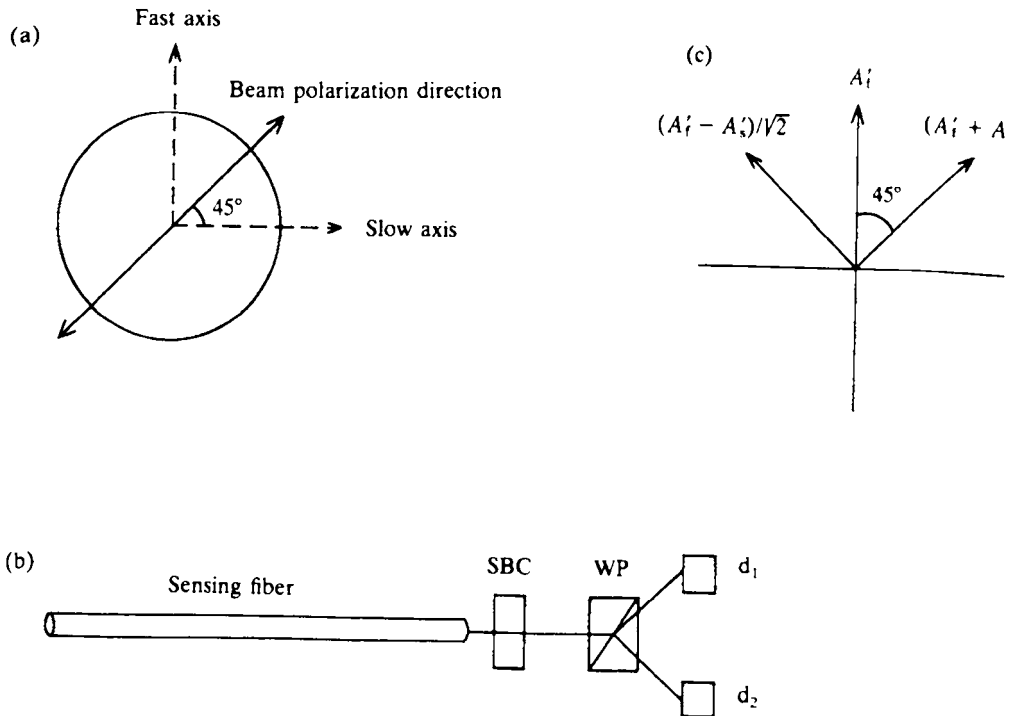


FIG. 10.11 In polarimetric sensors, light entering the sensing fiber has its polarization set at 45° to the 'fast' and 'slow' axes, as in (a). The emerging light is processed as shown in (b) by passing it through a Soleil-Babinet compensator (SBC) and then through a Wollaston prism (WP). The latter separates out the two orthogonally polarized beams which then fall onto detectors d_1 and d_2 . The amplitudes of the two output signals may be obtained by resolving the amplitudes as in (c).

modes to interfere, since they are orthogonally polarized. This is usually achieved with a combination of a Soleil–Babinet compensator and a Wollaston prism (Fig. 10.11). The former is a variable phase plate and it is set to introduce a phase difference of $\pi/2$ between the slow and fast modes. The amplitudes of the two beams can now be written as

$$A_r \exp(i\omega t) \quad \text{and} \quad A_s \exp[i(\omega t + \pi/2 + \Delta\phi_p)]$$

where $\Delta\phi_p$ is the phase difference between the beams before passing through the compensator. For simplicity, we write these two amplitudes as A'_r and A'_s .

A Wollaston prism is a device that splits a beam into two physically separate components having orthogonal polarization. In the present instance, the prism is set to resolve the emergent light into directions at 45° to the fast and slow directions. Thus the amplitudes of the two beams from the prism are given by $(A'_s + A'_r)/\sqrt{2}$ ($= A'_1$) and $(A'_s - A'_r)/\sqrt{2}$ ($= A'_2$) (Fig. 10.11). These signals are then allowed to fall onto two detectors d_1 and d_2 . Expanding the expressions for A'_s and A'_r , and assuming for simplicity that $A_s = A_r = A$, we obtain

$$A'_1 = \frac{A}{\sqrt{2}} \left\{ \exp(i\omega t) + \exp \left[i \left(\omega t + \frac{\pi}{2} + \Delta\phi_p \right) \right] \right\}$$

The beam irradiance I_1 ($= A_1 \times A_1^*$) is then

$$I_1 = \frac{A^2}{2} \left\{ \exp(i\omega t) + \exp \left[i \left(\omega t + \frac{\pi}{2} + \Delta\phi_p \right) \right] \right\} \left\{ \exp(-i\omega t) + \exp \left[-i \left(\omega t + \frac{\pi}{2} + \Delta\phi_p \right) \right] \right\}$$

When multiplied out this simplifies to

$$I_1 = A^2 [1 + \sin(\Delta\phi_p)]$$

Similarly, we may show that for the other beam the irradiance is given by

$$I_2 = A^2 [1 - \sin(\Delta\phi_p)]$$

The interferometer output is then most conveniently processed as

$$\frac{I_1 - I_2}{I_1 + I_2} = \sin(\Delta\phi_p)$$

For small values of $\Delta\phi_p$, the output is then proportional to $\Delta\phi_p$. The interferometer will of course only respond to the effect of external forces on the intrinsic birefringence B of the fiber. If we put $B = n_x - n_y$, then we have

$$\Delta\phi_p = \frac{2\pi LB}{\lambda_0}$$

If X represents the quantity to be measured then we may write

$$\frac{d\Delta\phi_p}{dX} = \frac{2\pi}{\lambda_0} \left(B \frac{dL}{dX} + L \frac{dB}{dX} \right) \quad (10.13)$$

Here, the first term on the right-hand side represents the effect of length change, whilst the second represents the change in birefringence. By comparing eq. (10.13) with eq. (10.8),

we can see that the former term is much smaller than the corresponding term in a conventional two-fiber interferometer, the ratio being in fact B/n_1 or $(n_x - n_y)/n_1$. For example, taking a fiber with a beat length of 2 mm, $n_x - n_y \sim 5 \times 10^{-4}$ (see eq. 8.48) and hence $(n_x - n_y)/n_1 \sim 3 \times 10^{-4}$. For fibers with a longer beat length, the factor will be correspondingly smaller. Thus we are justified in neglecting the first term on the right-hand side of eq. (10.13) in comparison with the second.

We saw earlier (section 8.7.5.1) that there are basically two types of high birefringence fiber: elliptically cored and internally stressed. The former tend to be relatively insensitive to pressure, axial stress and temperature, since none of these parameters can change the core ellipticity to any extent. Internally stressed fibers are produced by deliberately incorporating elements into the cladding which have different thermal expansion coefficients, thus producing an internal stress when they cool down during manufacture. Thus the internal stress, and hence B , is expected to be temperature dependent. Equally, axial strain affects the anisotropic stress within the fiber, again affecting B . External pressure, on the other hand, being isotropic with respect to the anisotropic strain, has little effect on B . Obviously the sensitivities to external forces are very dependent on the exact type of fiber used. However, for typical internally stressed fibers these are as follows: temperature, $3 \text{ rad K}^{-1} \text{ m}^{-1}$; pressure, $10^{-8} \text{ rad Pa}^{-1} \text{ m}^{-1}$; and strain, 10^6 rad .

It may seem perverse to develop sensors with deliberately reduced sensitivities. However, we have seen earlier that the high sensitivity of the conventional two-fiber interferometer, particularly with regard to temperature, can be something of an embarrassment. One interesting possibility is that of a remote polarimetric sensor. This consists of two parts: (a) a 'down' lead fiber and (b) a sensing fiber. Both fibers are of identical types except that they are fusion spliced together so that their fast and slow directions are at 45° (Fig. 10.12). The end of the sensing fiber is silvered to reflect light back down the fiber. In operation, linearly polarized light is launched into only one of the two fiber modes (the 'input' mode). When the light reaches the splice, energy will be launched into both polarization modes of the sensing fiber. After traversing the sensing fiber twice, the light will arrive back at the splice, and will then in general be coupled into the 'unused' mode of the input fiber (the 'output' mode). Back at the launch end, the light in this mode can be separated out and monitored. Any change in the differential path length in the sensing fiber will give rise to a change in the energy launched into the output mode.

10.1.4.4 Polarization rotation sensors

All the single mode fiber sensors considered so far have relied on the quantity to be measured affecting the phase of a mode within a fiber. Here, we consider the possibility of altering the angle of polarization of a mode. In section 3.7.1 we saw that such a rotation can take place in a medium in the presence of a magnetic field (Faraday rotation). Obviously, in the light of the discussion in section 8.7.5.1 we require a fiber that is as isotropic as possible, since any anisotropies can cause coupling between the orthogonally polarized modes. Unfortunately the Faraday effect is comparatively weak, and it is only possible to measure relatively large magnetic fields. One area where such a sensor has proved valuable is in the monitoring of large currents in electricity generating stations.

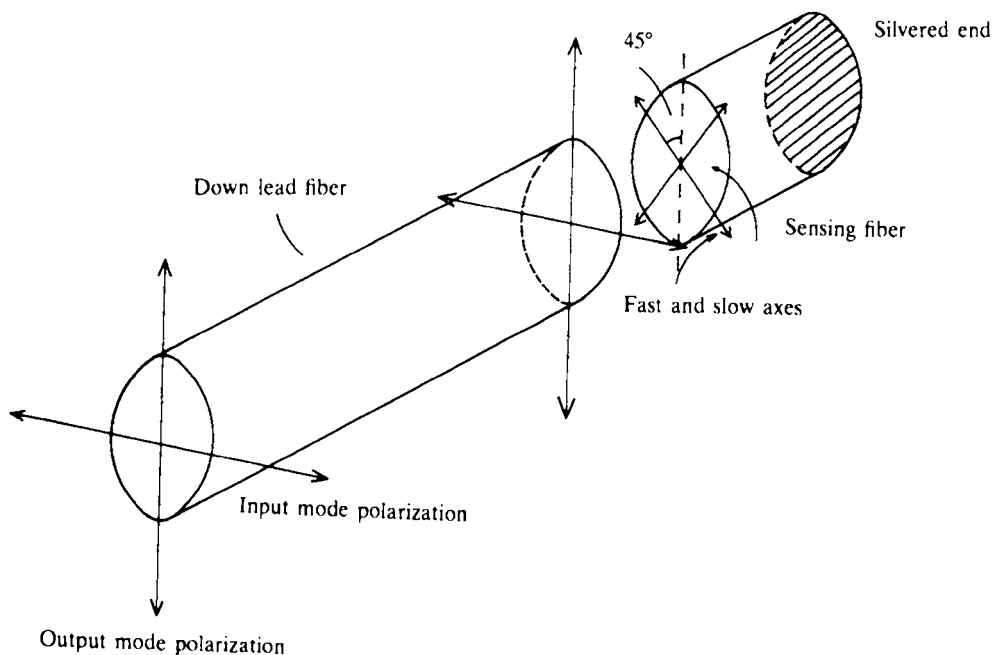


FIG. 10.12 Remote polarimetric sensor. The separation shown here between the down lead fiber and the sensing fiber is for pictorial convenience; normally the two are fused together.

Suppose we consider a single turn of fiber of radius r which passes round a wire carrying a current I A (Fig. 10.13). Ampere's circuital theorem gives $\mathcal{H} \times 2\pi r = I$, and hence $B = \mu_0 \mu_r I / 2\pi r$. The amount of polarization rotation is, from eq. (3.18), given by $\theta_r = VB2\pi r$; therefore, we have $\theta_r = \mu_0 \mu_r VI$, where μ_r is the relative permeability of the fiber and V is its Verdet constant. For n turns of fiber, the amount of rotation will be n times greater. Thus

$$\theta_r = \mu_0 \mu_r n VI \quad (10.14)$$

It will be noticed that this result is independent of r , and indeed in general θ_r is independent of the size or shape of the loop and the position of the conductor within the loop. This is a useful result, since it indicates that the device will be insensitive to any vibrations. As stated above the sensitivity of this device is low, and, as Example 10.5 shows, resolutions of some tens of amperes are all that can be expected.

EXAMPLE 10.5 Faraday rotation sensor sensitivity

For silica, $V = 4 \text{ rad m}^{-1} \text{ T}^{-1}$ and $\mu_r = 1$; hence for a 10 turn fiber coil and with $I = 30 \text{ A}$ the polarization rotation given by eq. (10.14) is

$$\theta_r = 4\pi \times 10^{-7} \times 10 \times 4 \times 30 \times \pi / 180^\circ = 0.09^\circ$$

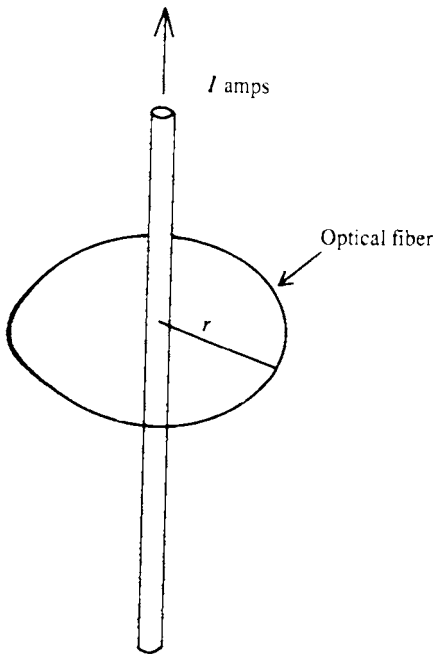


FIG. 10.13 Geometry for investigating the Faraday rotation induced in an optical fiber passing around a current-carrying conductor.

Normally, a resolution of about 0.1° in the polarization angle is possible, which is just the magnitude obtained here with a 30 A current.

In this section on optical fiber sensors, we have attempted to give an overall view of the many different types of sensor that have been proposed and of the basic physical principles behind them. Inevitably, in view of the limited space available, a comprehensive survey has not been possible. For a more detailed discussion, the interested reader may consult some of the many review articles and books that have been written (ref. 10.10). It should perhaps be emphasized that as yet few of these sensors have achieved widespread commercial success.

We move now to an area of application for optical fibers that has been making unspectacular but steady progress for many years, that is the use of bundles of fibers either for illumination or for imaging purposes.

10.2

Light-guiding fibers

Apart from their use in fiber optic communications, transparent optical fibers have been used for many years in a wide range of light-guiding applications in science, medicine and industry (ref. 10.11). Polished acrylic rods were first used for illumination purposes in the late

1920s. Most current applications, however, have arisen following the dramatic reduction in fiber losses resulting from Van Heel's suggestion that the fiber be provided with a cladding to reduce surface losses at each internal reflection within the fiber (ref. 10.12). Light propagates along these fibers in a zigzag manner in precisely the same way as we have described in section 8.3.

The commonest light-guiding fibers consist of a core and cladding as in communications fibers. The transmission distances involved are quite short, however, so there is not the same importance attached to having ultralow loss fibers. Usually, an all-glass (i.e. borosilicate) type of fiber is used with losses of the order of 500 dB km^{-1} . There is a sufficiently wide range of compatible glasses to enable fibers with a wide range of numerical apertures to be fabricated. The numerical aperture (NA) is given by eq. (8.23), although in the present case the amount of light which can be accepted by the fiber through the generation of skew rays is often important. With a Lambertian source, the light-gathering efficiency may be about 50% higher than that predicted by eq. (8.23), which, of course, applies only to meridional rays.

There are very few applications of single optical fibers (i.e. *monofibers*), but a brief consideration of these reveals the properties of fiber light guides. In general, the fiber can be regarded as an optical system which uses the light entering it to produce an output with the same solid angle but with reduced irradiance. Thus, a simple application is the facility of having a light source remote from, for example, a hazardous environment. Similarly, a fiber may be used to collect light from only part of a source in terms of angular or spatial dimensions. Monofibers can be used in this type of application subject to limitations caused by the dimensions of the fiber and the requirements for flexibility. The allowed bending radius at which a fiber breaks depends on the fiber diameter, and as a rough guide a fiber of $50 \mu\text{m}$ diameter will break at a bending radius of about 1 mm, while a fiber of 2 mm diameter may be taken as being non-flexible. If the transfer of light can be achieved without a requirement for flexibility, then larger diameter fibers can be used. Fibers up to 10 mm diameter can be bent, if required, by softening them in a flame, with negligible effect on their transmission characteristics. Such fibers are used for illuminating otherwise inaccessible regions where space is at a premium. They may, for example, be incorporated into the wall of a vessel to illuminate its inside, as illustrated in Fig. 10.14. If there is a requirement for the fiber to be flexible, either to allow for relative movement between the source and the point of application or because the path between these is not a straight line, then either small diameter monofibers or multifiber bundles may be used. It is quite straightforward to draw and assemble many thin fibers with diameters of about $50 \mu\text{m}$ into a bundle or *light guide*. In these devices, the ends of the fibers are bonded together leaving them free to flex over most of their length. The flexible portion is protected by, for example, a plastic sheath, which is bonded to a ferrule at each end of the bundle. Because of the cladding, each individual fiber behaves independently of the other fibers in the bundle and so there is no cross-talk.

The fibers are bonded together at their ends in a close-packed hexagonal array as illustrated in Fig. 10.15. Only the light entering the fiber cores is transmitted. A simple calculation based on Fig. 10.15 shows that in such an array the core packing fraction (i.e. the area occupied by the cores as a fraction of the total bundle cross-sectional area) is given by (see

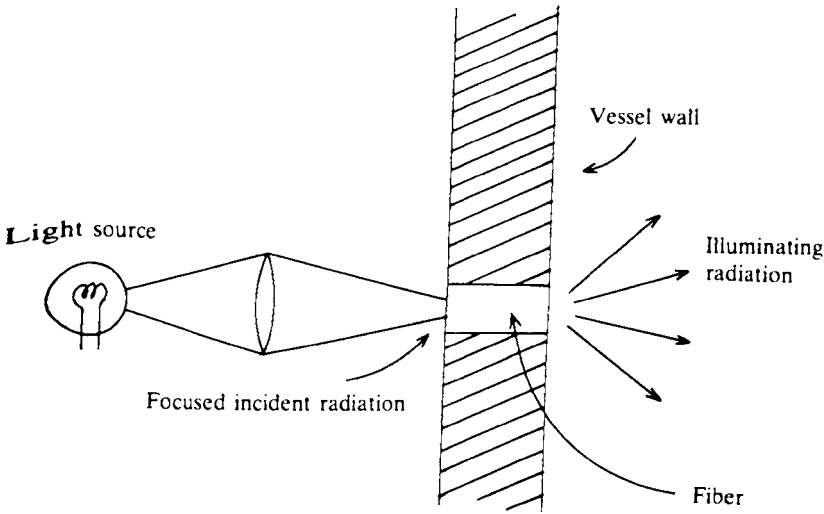


FIG. 10.14 Illumination of the inside of a vessel using a single fiber.

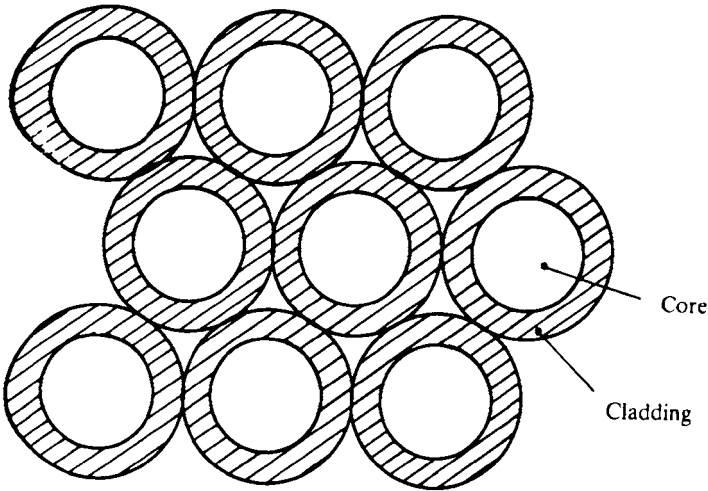


FIG. 10.15 Close-packed hexagonal array of fibers.

Problem 10.6)

$$F = \frac{\pi}{2\sqrt{3}} \frac{d^2}{D^2} \approx \frac{0.9d^2}{D^2} \quad (10.15)$$

where d is the core diameter and D the total fiber diameter. Typically, $F \approx 0.7$ indicating that the light guide has about 0.7 of the total transmission of the individual fibers comprising it.

In addition to simply transferring light between two points, light guides are used for shape converting and beam splitting. In shape converting, the ends of the bundle are bonded into

different shapes to match different input and output conditions. A typical application is the collection of light from a spectral source and its guidance to the slit of a spectrometer. Similarly, the output end of a light guide could be shaped to illuminate a single line of script in a document scanner.

In beam splitting guides, the total number of fibers is simply divided among several inputs or outputs (or both). Such beam splitting guides have been used for reading computer punched cards, while another common use is in illuminated signs such as those on motorways. Here a fiber bundle illuminated with a single light source is spread out to the appropriate parts of a matrix display (Fig. 10.16). Different display patterns can be formed by using different fiber bundles, each one being illuminated separately. To ensure that the angular distribution of the emerging light is uniform irrespective of which bundle is being used, a large diameter monofiber is situated at each matrix position to act as a 'mixer' (Fig. 10.17). Finally, a collimating lens reduces the divergence of the emerging beam, since a wide angle of view is not required, whereas it is essential to have as high a brightness as possible for the display to be visible in daylight conditions. Beam splitting guides are widely used as Y-guides in

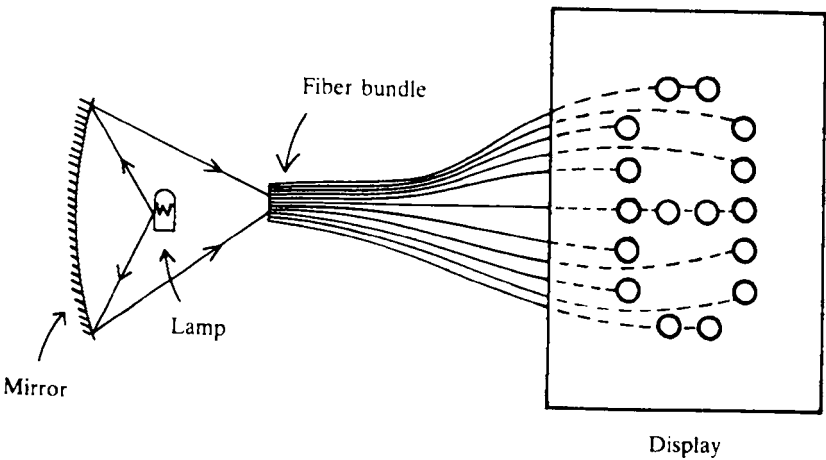


FIG. 10.16 Illumination of the legend on a motorway sign using a fiber bundle beam splitter.

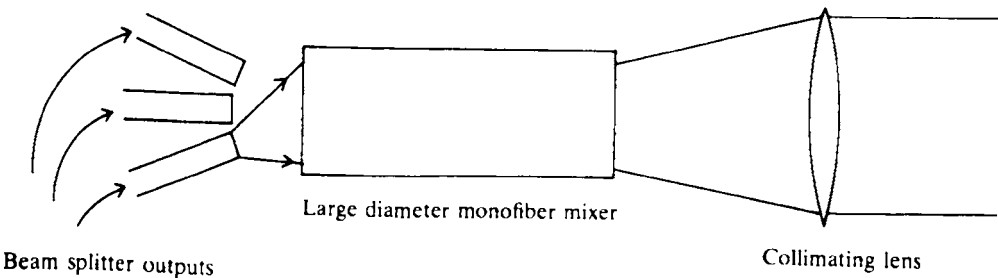


FIG. 10.17 Use of a monofiber to insure an even angular output from a number of beam splitter inputs for a motorway sign.

which the light traverses the guide in two directions. As shown in Fig. 10.2(a), light travels along one of the branches to the output end where it is reflected back into the other branch and guided to a detector. The detector signal then provides information about the conditions which exist in the vicinity of the common end. This system has been used for reflectance measurements, counting objects and proximity sensing, as we discussed in section 10.1.1.

10.2.1 Coherent bundles

Coherent bundles are assemblies of optical fibers in which the fibers are positioned accurately within the bundle such that spatial information can be transmitted. This is in marked contrast to the light guides discussed earlier where there is no correlation between the fiber positions at the two ends of the bundle.

In coherent bundles it is most important that there is no cross-talk between the constituent fibers, as this would degrade the information being transmitted. Cross-talk is prevented by making the cladding of the fibers sufficiently thick that light cannot leak from one fiber into neighbouring fibers. The output of each fiber will then depend solely on the light entering its input end face (within its acceptance angle) and the transmittance of the fiber. Hence, when an image is formed on the end of a coherent bundle the light emerging from each fiber is simply related to the irradiance distribution in the portion of the light illuminating its input face. Now, owing to the integrating nature of the individual fibers their outputs are of uniform irradiance so that no detail finer than the fiber diameter will appear in the transferred image.

The information transfer capacity is controlled by the resolution of the bundle, that is by the diameter of the individual fibers used. For this reason, the fibers used are of much smaller diameter than those used in light guides; diameters down to $5\text{ }\mu\text{m}$ are common. Such fibers are difficult to handle individually and are often produced as groups where the individual fibers are fused together into a multiple fiber, typically in a 6×6 square array. These multiple fibers are then stacked together to form a coherent bundle in their own right, containing up to a million individual fibers in lengths of a few metres with a resolution of about 50 line pairs per millimetre.

Flexible coherent bundles are fabricated by bonding the fibers together only at their extremities. Such bundles are quite widely used as *fiberscopes* for inspection purposes. An object lens is used to form a real image on the input face of the fiberscope. This image is then transferred to an eyepiece. Because the resolution of the fiberscope is limited by the diameter of the individual fibers a certain amount of defocusing of the image formed by the object lens may not be detectable by the observer (see Problem 10.7). Applications of fiberscopes range from the inspection of the inside of pipes and castings to the medical inspection of the human body – in medicine they are often referred to as endoscopes.

10.2.1.1 Rigid coherent bundles

The individual fibers in a bundle can be fused together by the application of heat and pressure to form a single mechanical unit. Physically, this behaves as a single piece of glass but, owing to the presence of the cladding on each fiber, it retains the optical properties of a coherent fiber bundle. In such bundles, the air gaps are totally removed in the fusing process so

that the core packing fraction is simply the square of the ratio of the core diameter to the total diameter. This type of bundle is often made from very small diameter fibers ($\leq 5 \mu\text{m}$), and because the cladding has to be thick enough to avoid cross-talk the packing fraction may be as low as 0.5, although packing fractions as high as 0.9 can also be achieved.

There are basically two categories of rigid coherent bundle. These are the image conduit, which has a length much greater than its diameter, and the fused plate, which has a length less than its diameter. Indeed, plates with diameters up to 200 mm are available.

During the manufacturing process, the fibers can be distorted around foreign particles present in the bundle and also any air not expelled may form bubbles, which again distort the fibers. These distortions allow light to escape from the affected fibers, thereby reducing their transmission relative to their neighbours with a consequent reduction in image quality.

Another consequence of the manufacturing process is that, as mentioned above, the bundle becomes effectively a single piece of glass so that stray light becomes much more important. In general, the light which falls on the end face of the cladding of a single fiber is not accepted by the fiber core but may be transmitted by internal reflections at the cladding/air interface. In long, flexible fibers, many such reflections occur so that the losses of stray light

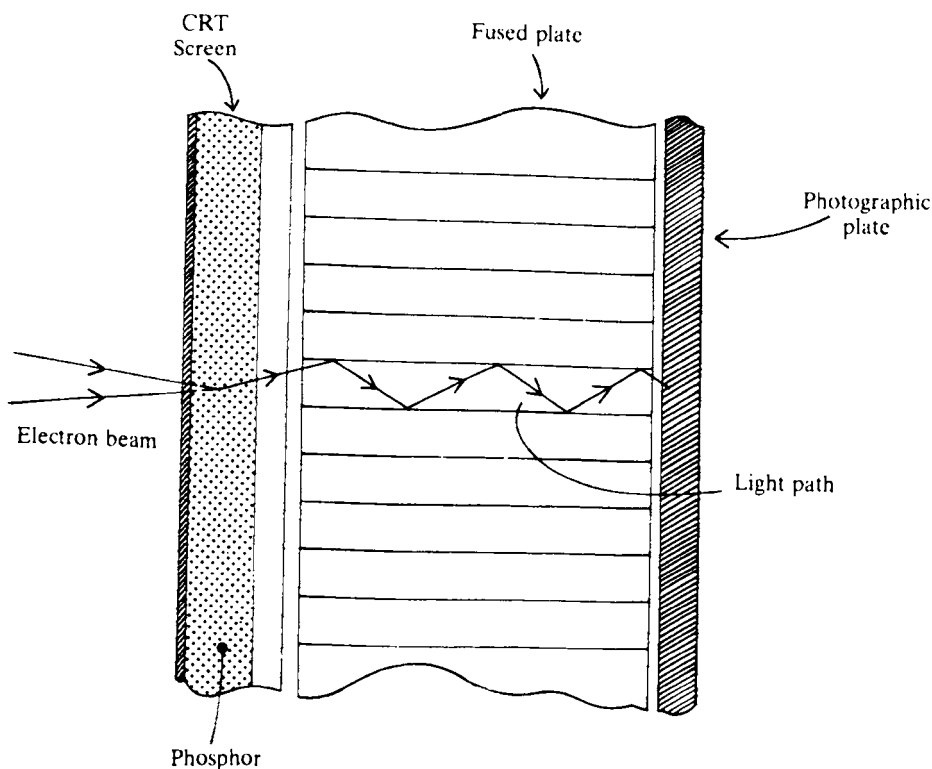


FIG. 10.18 Schematic cross-section of a faceplate allowing photographs to be taken from a CRT display.

are high and very little of it reaches the output face. In a rigid bundle, and more especially in fused plates, there are very few such reflections and a large fraction of the stray light may reach the output face. The effect of this is to increase the level of illumination uniformly over the transmitted image, thereby reducing the image contrast.

The most successful way of eliminating this unwanted light is to incorporate a second cladding of absorbing glass around the fiber – this is one form of what is known as *extramural absorption* (e.m.a.). The absorbing cladding does not affect the light accepted by the core but removes only stray light which may enter the inner cladding. Thus, the light reaching the output face of the bundle is predominantly that which forms the transmitted image.

While image conduits are not widely used, there are many applications for fused plates. Perhaps their major application is as face plates of electro-optic devices where the visual output is formed by the excitation of a phosphor. Fresnel reflection losses at the plate faces can be eliminated by the use of antireflection coatings so that up to 90% of the light emitted by the phosphor may be transmitted. This compares with a figure of less than 25% for lens systems.

A typical use of the fused plate is as a faceplate for photographing traces on a cathode ray tube. The plate must be vacuum-tight and capable of withstanding the changes of temperature which occur as the plate is coated with the phosphor. The trace excited by the electron beam is transmitted by the fused plate, as illustrated in Fig. 10.18, directly to the photographic film placed in contact with the outer face of the plate. The image intensifier described in section 7.3.4 also incorporates fused plates. In this case, the plates have a dual role, namely light transmission and image shaping.

PROBLEMS

- 10.1** Show that the displacement–output characteristic of the fiber displacement sensor shown in Fig. 10.1(c) is proportional to the factor $[\cos^{-1}\delta - \delta \sin(\cos^{-1}\delta)]/\pi$ where $\delta = 1 - (d/a)$, a being the radius of the expanded beam and d the displacement measured from the edge of the expanded beam. Assume a constant beam irradiance across the expanded beam and a straight-edged shutter.

Sketch this function over the range $d = 0$ to $2a$.

- 10.2** Calculate the fraction of light transferred from the emitting to the receiving fiber in the two-fiber version of the Fotonic sensor shown in Fig. 10.2(b) where two fibers of radius a and separation s are situated a distance z from a reflecting surface. (Hint: as implied in the text, consider the ‘coupling’ between the image of the emitting fiber and the receiving fiber. There are three regions to be considered: (a) $R < a + s$, (b) $s + 3a \geq R \geq s + a$, and (c) $R > s + a$, where $R = a + 2z \tan[\sin^{-1}(\text{NA})]$. In the second region, where the radiation cone partially overlaps the receiving fiber core, assume that the circular edge of the cone may be replaced by a straight edge; this enables the result of Problem 10.1 to be used.)

Sketch the coupling efficiency as a function of z for typical values of a and s .

- 10.3** It is proposed to use the amount of light transmitted between two fibers whose ends are offset longitudinally (see Fig. 8.34 and Problem 8.17) as an optical ‘dipstick’.

Investigate how the amount of light transmitted between the fibers varies with the refractive index and the absorption coefficient of the intervening medium.

If the sensor is lowered from air to water, by how much would you expect the coupled radiation to increase?

- 10.4 Derive the form of the output expected from a Mach–Zehnder interferometer when beam irradiances in the two arms are unequal. What advantage is there in this situation in using both possible outputs from the interferometer?
- 10.5 Use a simple meridional ray model (i.e. Fig. 8.6) to calculate the phase changes expected in a fiber of length L and radius a for a given mode when there is (a) a length change of δL and (b) a radius change of δa . Ignore the phase change on reflection. Hence show that for modes far from cut-off, the same fractional change in L will give rise to a much greater phase change than will the same fractional change in a .
- 10.6 Show that with a close-packed hexagonal bundle of fibers the ratio of core area to the total bundle cross-sectional area is given by $\pi d^2/(2\sqrt{3}D^2)$ where d is the core diameter of an individual fiber and D its total diameter.
- 10.7 The light from a point object is focused with unit magnification onto the end of a fiber whose core diameter is d . Show that the fiber may be moved backward and forward a distance $fd/2a$ about the point of focus without altering the total amount of light entering the fiber core. When fiber bundles are used for image transmission, this represents the extent to which the lens can be defocused and still give the same resolution in the transmitted image.

REFERENCES

- 10.1 J. Gowar, *Optical Communication Systems* (2nd edn), Prentice Hall International, Hemel Hempstead, 1993, section 9.7.
- 10.2 C. D. Butler and G. B. Hocker, 'Fibre optics strain gauge', *Appl. Opt.*, **17**, 2867, 1978.
- 10.3 B. Culshaw and J. Dakin (eds), *Optical Fiber Sensors: Systems and Applications*, Vol. 2, Artech House, Norwood, MA, 1989, Chapter 16.
- 10.4 D. A. Long, *Raman Spectroscopy*, McGraw-Hill, London, 1977.
- 10.5 J. P. Dakin (ed.), *The Distributed Fibre Optic Sensing Handbook*, Springer-Verlag, Berlin, 1990.
- 10.6 T. Kurashima, T. Horigouchi and M. Tateda, 'Distributed-temperature sensing using stimulated Brillouin scattering in optical silica fibres', *Opt. Lett.*, **15**, 1038, 1990.
- 10.7 E. Udd (ed.), *Fiber Optic Smart Structures*, John Wiley, New York, 1995.
- 10.8 R. S. Longhurst, *Geometrical and Physical Optics* (3rd edn), Longman, Harlow, 1973, Chapter 21.
- 10.9 J. C. D. Jones, P. A. Leilabody and D. A. Jackson, 'Monomode fiber-optic sensors: optical processing schemes for recovery of phase and polarization state information', *Int. J. Opt. Sens.*, **1**, 123, 1986.

- 10.10** (a) T. D. Giallorenzi, J. A. Bucaro, A. Dandridge, G. H. Sigen, J. H. Cole, S. C. Rashleigh and R. G. Priest, 'Optical fiber sensor technology', *IEEE J. Quantum Electron.*, **18**, 626, 1982.
- (b) D. A. Jackson and J. D. C. Jones, 'Fiber optic sensors', *Opt. Acta*, **33**, 1469, 1986.
- (c) B. Culshaw, *Optical Fiber Sensing and Signal Processing*, Peter Peregrinus, London, 1984.
- 10.11** (a) W. P. Siegmund, 'Fiber optics', in R. Kingslake (ed.) *Applied Optics and Optical Engineering*, Vol. 4, pp. 1-29, Academic Press, New York, 1967.
- (b) W. B. Allan, *Fibre Optics* (Engineering Design Guides 36), published for the Design Council, British Standards Institution and the Council for Engineering Institutions by Oxford University Press, Oxford, 1980.
- 10.12** A. C. S. Van Heel, 'Optical representation of images with the use of lenses or mirrors', *Ingenieur*, **65**, 25, 1953.

Answers to numerical problems

- 1.3** $5.3 \times 10^5 \text{ V m}^{-1}$
1.6 0.8 mm; 47 mm
1.7 $1.5 \times 10^{-5} \text{ m}$; for $\lambda_0 = 600 \text{ nm}$, 35 and 40; for $\lambda_0 = 500 \text{ nm}$, 42 and 48
1.8 $6.25 \times 10^{-8} \text{ m}$; $12.5 \times 10^{-8} \text{ m}$
1.9 Six orders; 409.5 nm; 468; 546 nm; 655.2 nm
1.10 0.77 mm; 15.4 mm
1.12 $4.08 \times 10^{-19} \text{ J}$ (2.55 eV); 487 nm
1.13 $5.80 \times 10^{-14} \text{ Hz}$; $2.79 \times 10^{-19} \text{ J}$ (1.74 eV)
2.3 $5 \times 10^{-11} \text{ m}$
2.4 $1.84 \times 10^5 \text{ s}^{-1}$
2.5 $6.02 \times 10^{-18} \text{ J}$; $2.41 \times 10^{-17} \text{ J}$; $5.42 \times 10^{-17} \text{ J}$; $4.82 \times 10^{-17} \text{ J}$; $7.27 \times 10^{16} \text{ Hz}$
2.7 $1.64 \times 10^{16} \text{ m}^{-3}$; $224 \Omega^{-1} \text{ m}^{-1}$; $80 \Omega^{-1} \text{ m}$
2.8 0.008 eV
2.9 0.127; 0.873
2.11 $8.1 \times 10^{11} \text{ m}^{-3}$
2.12 $3.5 \times 10^{-3} \text{ m}^2 \text{ s}^{-1}$; $1.25 \times 10^{-3} \text{ m}^2 \text{ s}^{-1}$; $4.18 \times 10^{-4} \text{ m}$; $2.5 \times 10^{-4} \text{ m}$
2.15 $1 \times 10^{23} \text{ m}^{-3}$
2.16 0.9 V; $4.88 \times 10^{-8} \text{ m}$; $3.68 \times 10^{-7} \text{ m}$
2.17 0.72 μA ; 99%
2.18 868 nm; 517.6 nm; 5.52 μm
3.2 Quartz 0.0162 mm; calcite 1.71 μm
3.3 8.4×10^{-5}
3.4 KDP 7.53 kV; KD*P 3.03 kV; ADP 9.2.1 kV; LiNbO₃ 0.74 kV
3.5 1400 V
3.7 1.45 W
3.9 0.4°
4.2 260 MHz
4.4 2.6 nits
4.5 $2\pi\text{AB}(0)$; $\pi\text{AB}(0)$
4.7 0.5
5.1 41 532 K; 69.2 μm
5.2 10.05 m^{-1} ; 0.366

5.3	1.39 m^{-1} ; 0.098 m^{-1}
5.4	$9.30 \times 10^{-29}\%$; $9.57 \times 10^{-17}\%$
5.5	$5.6 \times 10^8 \text{ m}^{-3}$
5.6	$5.19 \times 10^7 \text{ Hz}$; $1.28 \times 10^9 \text{ Hz}$
5.7	90.5% reflectance
5.8	0.64 m^{-1}
5.9	1 580 278; $3 \times 10^8 \text{ Hz}$; five
5.11	6.72 MW m^{-3}
5.12	4.22 kW
5.13	0.7
6.1	0.1 m; 0.21 K
6.2	474.6 m s^{-1}
6.4	$4 \times 10^7 \text{ Hz}$
6.5	0.316 mm
6.6	$4.05 \times 10^{11} \text{ W m}^{-2} \text{ sr}^{-1}$; $6.89 \times 10^{13} \text{ lm m}^{-2} \text{ sr}^{-1}$
6.7	3.33 ns; 667 ps; 13.4 ns; 17.5 fs; 0.44 ps; 2.05 ns
6.8	5 m; 122.45 m
6.9	17.59 MeV
7.1	10^{-10} W
7.2	48 μm
7.4	$5.5 \times 10^{-11} \text{ W}$
7.5	$1.7 \times 10^{-6} \text{ A V}^{-1} \text{ J}^{-1}$
7.7	$10^{12} \text{ photons s}^{-1}$
7.10	$7.26 \times 10^{-11} \text{ W}$; $2.44 \times 10^{-12} \text{ W}$
7.11	6.4 mA; 0.346 V
7.13	$y = 0.922$; $x = 0.427$
7.15	57 μm
8.1	48.8°
8.3	$\psi = 63.8^\circ$; $\delta = 64.7^\circ$
8.4	0.52%; $1.3 \times 10^{-21}\%$
8.6	2230 dB m^{-1}
8.8	9.2°; 1.441; 1263; yes to 6.9°
8.9	30 ns km^{-1}
8.10	0.68 μm ; 3.6 μm
8.11	150 ps km^{-1}
8.14	52.3 dB km^{-1}
8.15	$0.0294 \text{ dB km}^{-1}$, 2.29 μm
8.18	5.5 dB km^{-1}
8.19	About 10 MHz
8.23	About $4 \times 10^{-3} \text{ dB km}^{-1}$
9.1	0.391% (1 part in 256)
9.2	64 Mbps

9.3	6 km (range limited by absorption)
9.4	14.9 μW
9.7	0.923
9.9	S/N 221 (\equiv 23.4 dB); 318 Ω ; S/N 1387 (\equiv 31.4 dB)
9.11	21.5; 10.5 dB
9.12	10 nW
9.15	1.63 mm
9.16	0.86 V
10.3	1.77 times

Birefringence

Anisotropy in crystals is caused by the arrangement of atoms being different in different directions through the crystal. Consequently, the electric polarization (or dipole moment per unit volume) \mathbf{P} produced in a crystal by a given electric field \mathbf{E} depends on the direction of the field. It follows that the relative permittivity ϵ_r and the refractive index, which is given by $n = (\epsilon_r)^{1/2}$ (ref. A2.1, Chapter 11), are also direction dependent.

Because electromagnetic waves are transverse, the refractive index for a given direction of propagation of light through the crystal depends on the value of the electric vector \mathbf{E} , which is, of course, perpendicular to this direction. In fact, the electric vector oscillates in a plane perpendicular to the wave velocity direction and thus the velocity of the waves through the crystal will be dependent on the plane of polarization of the light. This phenomenon, called *birefringence*, occurs naturally in some materials, and it may be enhanced by the application of external forces such as those due to electric fields. It may also be induced in isotropic materials by the same means.

A2.1

Natural birefringence

We may appreciate how birefringence arises as follows. Let us consider placing a dielectric medium into a parallel plate capacitor to which a constant voltage is applied. The electric field across the medium polarizes it and dipoles appear throughout the material, lined up in the direction of the field. The resulting polarization can be shown to be related to the electric field and electric displacement, \mathbf{D} , by the equations (ref. A2.1, Chapter 3)

$$\begin{aligned}\mathbf{D} &= \epsilon_0 \mathbf{E} + \mathbf{P} \\ \mathbf{D} &= \epsilon_0 \epsilon_r \mathbf{E} = \epsilon \mathbf{E}\end{aligned}\tag{A2.1}$$

The ratio $\mathbf{P}/\epsilon_0 \mathbf{E}$ is called the electric *susceptibility*, χ , and we see from eqs (A2.1) that

$$\mathbf{P}/\epsilon_0 \mathbf{E} = (\epsilon_r - 1) = \chi\tag{A2.2}$$

In isotropic media (and cubic crystals), direction in the medium is not important, the vectors \mathbf{E} , \mathbf{D} and \mathbf{P} are all parallel, and ϵ_r , n and χ are scalar quantities. The speed of propagation of electromagnetic waves in such media is constant irrespective of the direction of propagation. This is not the case in birefringent materials. In general, in such materials, the polarization occurring in response to an applied electric field depends on both the magnitude and

direction of the field and, in addition, the induced polarization may be in a different direction from that of the field. Thus, for example, a field applied in the x direction will induce not only a polarization P_x parallel to itself, but possibly also polarizations P_y and P_z in the y and z directions. Thus, from eq. (A2.2) we may write

$$\begin{aligned} P_x &= \epsilon_0(\chi_{11}\mathcal{E}_x + \chi_{12}\mathcal{E}_y + \chi_{13}\mathcal{E}_z) \\ P_y &= \epsilon_0(\chi_{21}\mathcal{E}_x + \chi_{22}\mathcal{E}_y + \chi_{23}\mathcal{E}_z) \\ P_z &= \epsilon_0(\chi_{31}\mathcal{E}_x + \chi_{32}\mathcal{E}_y + \chi_{33}\mathcal{E}_z) \end{aligned} \quad (\text{A2.3})$$

where the capital letters represent the complex amplitudes of the corresponding time-varying quantities. The 3×3 matrix array χ_{ij} is called the electric susceptibility tensor. Alternatively, we can use the electric permittivity tensor ϵ_{ij} where $D_i = \epsilon_{ij}\mathcal{E}_j$ to describe the dielectric response of the crystal to an applied field. These tensors are also 3×3 matrix arrays.

The magnitude of the χ_{ij} (and ϵ_{ij}) coefficients depends on the choice of the x , y and z coordinate axes relative to the crystal structure. It turns out always to be possible to choose x , y and z so that the off-diagonal elements vanish, giving

$$\begin{aligned} P_x &= \epsilon_0\chi_{11}\mathcal{E}_x \\ P_y &= \epsilon_0\chi_{22}\mathcal{E}_y \\ P_z &= \epsilon_0\chi_{33}\mathcal{E}_z \end{aligned} \quad (\text{A2.4})$$

and similarly

$$\begin{aligned} D_x &= \epsilon_{11}\mathcal{E}_x \\ D_y &= \epsilon_{22}\mathcal{E}_y \\ D_z &= \epsilon_{33}\mathcal{E}_z \end{aligned} \quad (\text{A2.5})$$

where from eqs (A2.1) and (A2.4) we see that $\epsilon_{11} = \epsilon_0(1 + \chi_{11})$ etc. These directions are called the *principal axes* of the crystal and the corresponding diagonal terms ϵ_{11} , ϵ_{22} and ϵ_{33} the *principal permittivities*.

An electric field parallel to a principal axis will only produce a parallel electric polarization but in general the three permittivities are different. This variation in permittivity and the corresponding variation in refractive index and hence in wave velocity is the origin of birefringence. In general, anisotropic crystals are thus characterized by the three *principal refractive indices* n_x , n_y and n_z , where the subscripts relate to the direction of polarization of the electromagnetic waves and *not* to their direction of propagation, and where $n_x = (\epsilon_{11}/\epsilon_0)^{1/2} = (1 + \chi_{11})^{1/2}$ etc.

Thus consider, for example, a wave propagating in the z direction. If its electric field, that is its plane of polarization, is parallel to the x direction, it will induce only polarization P_x according to eqs (A2.4) and (A2.5) and will experience the permittivity ϵ_{11} , that is the refractive index n_x . If, on the other hand, the wave has its plane of polarization in the y direction, it will experience the refractive index n_y . Any unpolarized wave propagating in the z direction can be resolved into two components with polarizations parallel to the x and y directions. These two components will thus travel through the crystal with different velocities,

and hence get progressively out of phase, so that the resultant at some point z will, in general, represent elliptically polarized light. The exact nature of the light depends on the phase difference and the relative amplitudes of the two components.

Anisotropic crystals are divided into two classes. The first class, which includes calcite (CaCO_3), are characterized by a principal axis (the z axis by convention) along which the permittivity is ϵ_{33} , but perpendicular to which the permittivity is independent of direction (i.e. $\epsilon_{11} = \epsilon_{22}$). Thus, such crystals only have two principal refractive indices. The z direction is unique in that the velocity of propagation is independent of the wave polarization; it is termed the *optic axis*. Such materials are said to be uniaxial. The difference between n_x ($=n_y$) and n_z is a measure of the birefringence and is said to be positive or negative depending on whether n_z is greater or smaller than n_x . Thus calcite is a negative uniaxial crystal, while quartz, for example, is a positive uniaxial crystal.

Crystals of lower symmetry possess two optic axes and are said to be *biaxial*. These axes do not coincide with any of the principal axes and consequently n_x , n_y and n_z are all different.

If Maxwell's electromagnetic theory is applied to an isotropic medium, it can be shown that light of any polarization can propagate in any direction. In the case of anisotropic materials, however, this is not the case and it turns out (ref. A2.3) that as a result of birefringence only two states of polarization can propagate for any crystal direction. These planes are perpendicular to each other and hence, except when propagation is parallel to an optic axis, the two components experience different refractive indices and travel at different speeds.

The values of refractive index and the states of polarization for a particular direction of propagation can be determined from a three-dimensional representation of the refractive index known as the *index ellipsoid* (or *optical indicatrix*) (ref. A2.2). The form of this can be obtained as follows: the energy density W in a dielectric is given by

$$W = \frac{1}{2} \mathbf{D} \cdot \mathbf{E} \quad (\text{A2.6})$$

Thus using eq. (A2.5), we have

$$W = \frac{1}{2} \left(\frac{D_x^2}{\epsilon_{11}} + \frac{D_y^2}{\epsilon_{22}} + \frac{D_z^2}{\epsilon_{33}} \right)$$

A constant energy surface then takes the form of an ellipsoid in a three-dimensional space where the dielectric displacement is the coordinate.

Dividing both sides by W and writing $D_x^2/2\epsilon_0 W$ as x^2 etc. leads to the expression

$$\frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1 \quad (\text{A2.7})$$

where $n_x^2 = \epsilon_{11}/\epsilon_0$ etc.

This equation represents an ellipsoid with semiaxes n_x , n_y and n_z . For a uniaxial crystal, $n_x = n_y \neq n_z$ and the index ellipsoid, which has circular symmetry about the z axis, is as shown in Fig. A2.1.

It is usual to refer to n_x ($=n_y$) as the *ordinary* refractive index n_o and to n_z as the

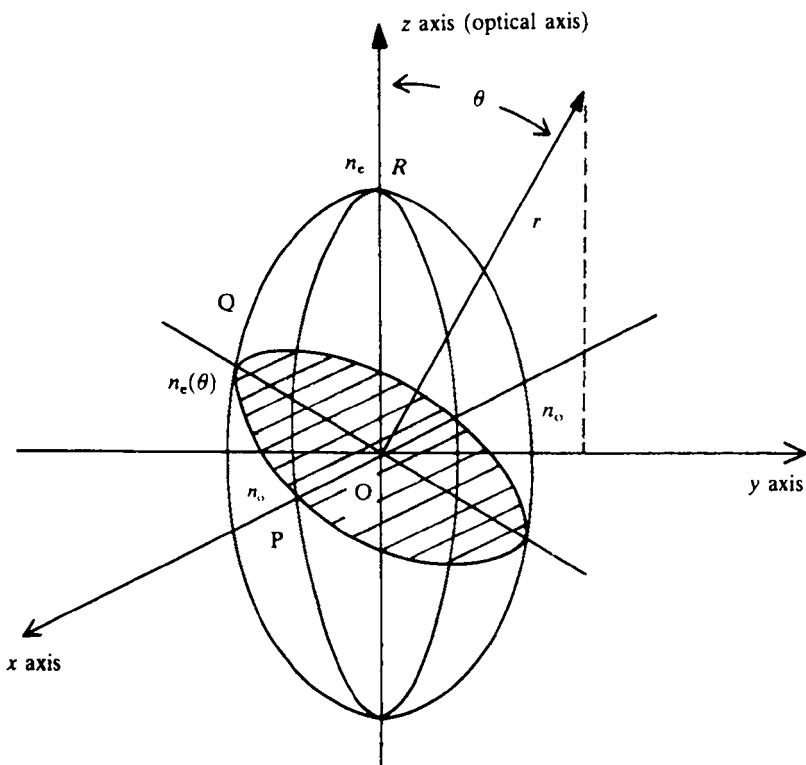


FIG. A2.1 Refractive index ellipsoid or optical indicatrix for a uniaxial crystal. Waves with polarizations parallel to the z axis experience the refractive index n_e , while those polarized parallel to the x and y axes experience the refractive index n_o . For propagation of a wave in the general direction \mathbf{r} , there are two allowed directions of polarization: parallel to OP , with index n_o , and parallel to OQ , with index $n_e(\theta)$.

extraordinary refractive index n_e , in which case we may write

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = 1 \quad (\text{A2.7a})$$

Let us consider light propagating in a direction \mathbf{r} , at an angle θ to the optic (z) axis (Fig. A2.1). Because of the circular symmetry mentioned above, we can choose, without loss of generality, that the y axis should coincide with the projection of \mathbf{r} on the xy plane. The plane normal to \mathbf{r} intersects the ellipsoid in the shaded ellipse shown. The two allowed directions of polarization are parallel to the axes of this ellipse and thus correspond to OP and OQ ; they are thus perpendicular to \mathbf{r} as well as to each other. The two waves polarized along these directions have refractive indices given by $OP \equiv n_o$ and $OQ \equiv n_e(\theta)$ (the extraordinary wave). In the case of the extraordinary wave, the plane of polarization varies with θ as does the refractive index. We can determine the relationship between $n_e(\theta)$ and n_e , n_o with the aid of Fig. A2.2, which shows the intersection of the index ellipsoid with the yz plane. From the diagram,

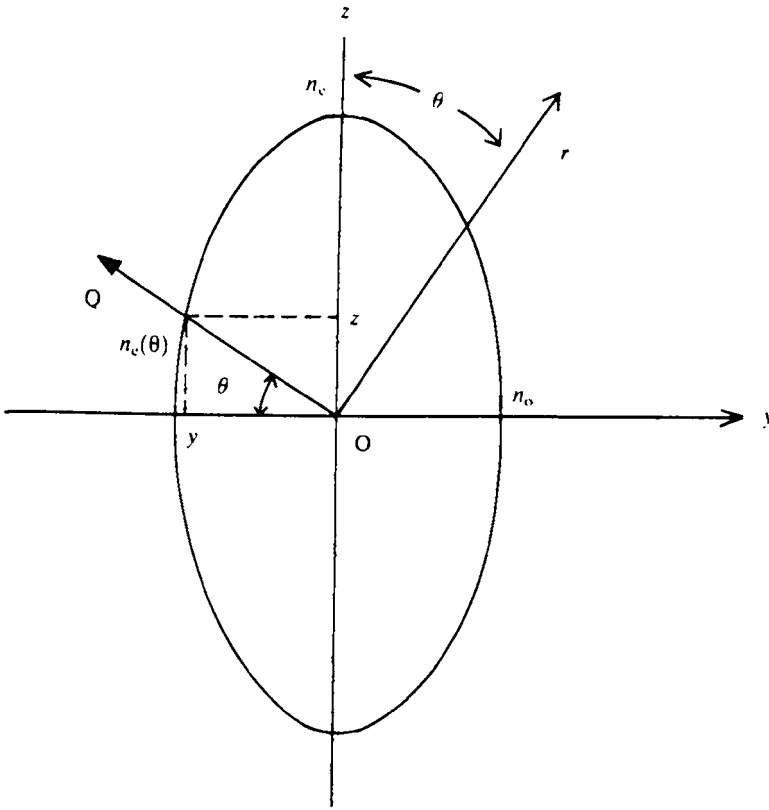


FIG. A2.2 Intersection of the index ellipsoid with the yz plane. $OQ = n_e(\theta)$ represents the refractive index of the extraordinary wave propagating in the r direction.

we see that

$$n_e^2(\theta) = z^2 + y^2$$

and

$$z = n_e(\theta) \sin \theta$$

Thus, substituting these equations into the equation for the ellipse shown, namely

$$\frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = 1$$

gives

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2 \theta}{n_o^2} + \frac{\sin^2 \theta}{n_e^2}$$

Thus for $\theta = 0^\circ$, that is propagation along the optic axis, $n_e(0^\circ) = n_o$, while for $\theta = 90^\circ$,

$n_e(90^\circ) = n_e$. The two polarizations which can be propagated correspond to the maximum and minimum refractive indices given by the index ellipsoid (in the case of positive crystals $n_e > n_o$, while in the case of negative crystals $n_e < n_o$). For propagation parallel to the optic axis (i.e. the z direction), there is no birefringence as the section of the ellipsoid perpendicular to this direction is a circle. For propagation perpendicular to the optic axis, for example the x direction, the birefringence will be a maximum, the permitted polarizations will be parallel to the y axis with refractive index n_o and parallel to the z axis with refractive index n_e .

A2.2

Induced birefringence

When an electric field is applied to an anisotropic crystal, new principal axes may be introduced (e.g. in KDP) while birefringence may be induced in naturally isotropic materials such as GaAs. The electric field changes the shape of the refractive index ellipsoid as a result of changing the permittivity tensor ϵ_{ij} . After deformation, the axes of the ellipsoid are no longer the original principal axes and the ellipsoid must be represented by an equations of the form

$$\left(\frac{1}{n^2}\right)_1 x^2 + \left(\frac{1}{n^2}\right)_2 y^2 + \left(\frac{1}{n^2}\right)_3 z^2 + 2\left(\frac{1}{n^2}\right)_4 yz + 2\left(\frac{1}{n^2}\right)_5 xz + 2\left(\frac{1}{n^2}\right)_6 xy = 1 \quad (\text{A2.8})$$

The applied electric field causes a change in the refractive indices of the ordinary and extraordinary refractive indices. This change can be written as

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^3 r_{ij} \mathcal{E}_j \quad (\text{A2.9})$$

where $i = 1, 2, \dots, 6$, $j = 1, 2, 3$, and r_{ij} is the electro-optic tensor with 6×3 elements. For most materials, however, the symmetry of the crystal reduces the 18 elements of this tensor significantly. We have indicated the values of some of these electro-optic coefficients in Table 3.1 for a representative selection of optical materials. We see that the Pockels effect discussed in Chapter 3 will thus be different in the various crystals. In general, crystal class and symmetry (ref. A2.3) are important and a universal response cannot be predicted. Let us consider as an example crystals similar to KDP. In this class of materials, the only non-zero elements in the electro-optic tensor are r_{41} , r_{52} ($= r_{41}$) and r_{63} . In an arbitrarily directed field, the index ellipsoid becomes

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} + 2r_{41}\mathcal{E}_x yz + 2r_{41}\mathcal{E}_y xz + 2r_{63}\mathcal{E}_z xy = 1 \quad (\text{A2.10})$$

We note that the first three terms are field independent and hence involve n_o and n_e , and that the field induces cross-terms yz etc. into the equation for the index ellipsoid so that the major axes of the ellipsoid are no longer parallel to the x , y and z axes as mentioned above. We must now therefore find the magnitudes and directions of the new axes, in the presence of the field, so that we may determine its effect on the propagation of electromagnetic waves. It is usual to restrict the field to the z direction (the original optical axis) so that $\mathcal{E}_x = \mathcal{E}_y = 0$

and eq. (A2.10) becomes

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} + 2r_{63}\mathcal{E}_z xy = 1 \quad (\text{A2.11})$$

The problem now is to find a new coordinate system x', y', z' in which eq. (A2.11) contains no cross-term and takes up the form of eq. (A2.7), that is

$$\frac{x'^2}{n_{x'}^2} + \frac{y'^2}{n_{y'}^2} + \frac{z'^2}{n_{z'}^2} = 1 \quad (\text{A2.12})$$

The lengths of the major axes of the index ellipsoid are thus $2n_{x'}$, $2n_{y'}$, $2n_{z'}$ and will, in general, depend on the applied field. By inspection of eqs (A2.11) and (A2.12), the z factor is unchanged if z is parallel to z' and hence $n_{z'} = n_e$, and also eq. (A2.11) is symmetric in x and y so the new axes x', y' bear similar relations to the former axes x and y ; they are, in fact, rotated through 45° as shown in Fig. A2.3. The relationships between the two sets of axes are then

$$\begin{aligned} x &= x' \cos 45^\circ + y' \sin 45^\circ \\ y &= -x' \sin 45^\circ + y' \cos 45^\circ \end{aligned} \quad (\text{A2.13})$$

Substituting eqs (A2.13) into eq. (A2.11) then gives

$$x'^2 \left(\frac{1}{n_o^2} - r_{63}\mathcal{E}_z \right) + y'^2 \left(\frac{1}{n_o^2} + r_{63}\mathcal{E}_z \right) + \frac{z^2}{n_e^2} = 1 \quad (\text{A2.14})$$

Equation (A2.14) shows that x', y', z' are the new principal axes and that the field has rotated the ellipse and changed the lengths of the major axes. From eqs (A2.12) and (A2.14), the

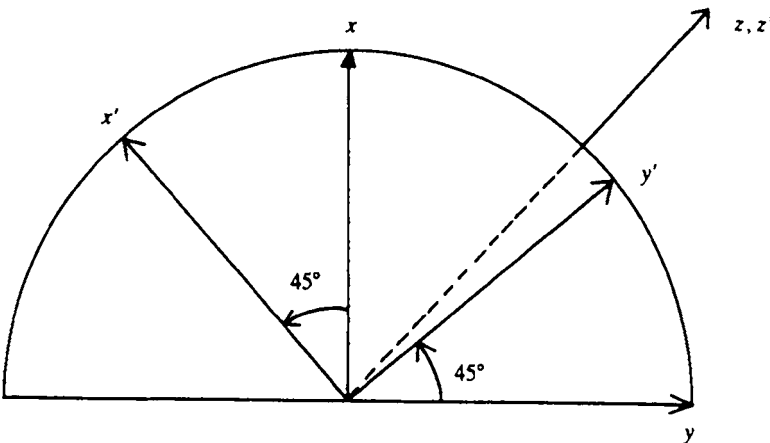


FIG. A2.3 Rotation of the x and y axes into the x' and y' axes, when crystals such as KDP are subjected to an electric field along the z axis, in which case z and z' are parallel.

length of the major axis in the x' direction, $2n_{x'}$, is given by

$$\frac{1}{n_{x'}^2} = \frac{1}{n_o^2} - r_{63} \mathcal{E}_z$$

Hence the change in refractive index may be evaluated by noting that

$$\frac{1}{n_{x'}^2} - \frac{1}{n_o^2} = \frac{(n_o - n_{x'})(n_o + n_{x'})}{n_{x'}^2 n_o^2} = \frac{2(n_o - n_{x'})}{n_o^3}$$

assuming that the change is small so that $n_{x'} \approx n_o$. Therefore

$$n_{x'} = n_o + \frac{n_o^3}{2} r_{63} \mathcal{E}_z \quad (\text{A2.15})$$

and similarly

$$n_{y'} = n_o - \frac{n_o^3}{2} r_{63} \mathcal{E}_z \quad (\text{A2.16})$$

and, as mentioned above, $n_z = n_e$.

Similar relationships can be derived for other classes of crystal and in general $n = \pm \frac{1}{2} n_o^3 r \mathcal{E}$, but the matrix element r varies according to the crystal class (see Table 3.1).

REFERENCES

- A2.1** P. Lorrain and D. Corson, *Electromagnetic Fields and Waves* (3rd edn), Freeman, New York, 1988, Chapter 9.
- A2.2** (a) M. Born and E. Wolf, *Principles of Optics*, Macmillan, New York, 1964.
 (b) A. Yariv, *Quantum Electronics* (2nd edn), John Wiley, New York, 1975.
 (c) A. Yariv and P. Yeh, *Optical Waves in Crystals*, John Wiley, New York, 1983.
- A2.3** See, for example, ref. 3.3a or N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart & Winston, New York, 1976, Chapter 4.

Limitations on LED frequency response due to carrier diffusion and recombination

We consider a relatively low level injection of electrons into a semi-infinite piece of p-type material with unit cross-sectional area. The steady state diffusion equation (eq. 2.41) for the excess minority carrier density at a point x , $\Delta n(x)$, may be written

$$\frac{\partial}{\partial t} \Delta n(x) = -\frac{\Delta n(x)}{\tau_e} + D_e \frac{\partial^2}{\partial x^2} \Delta n(x) \quad (\text{A3.1})$$

We suppose that the excess population has a time-varying component and we write

$$\Delta n(x) = \Delta n^0(x) + \Delta n^1(x) \exp(2\pi i f t) \quad (\text{A3.2})$$

Substituting for $\Delta n(x)$ in eq. (A3.1) and separating out the time-dependent and time-independent parts, we obtain

$$\frac{\partial^2}{\partial x^2} \Delta n^0(x) - \frac{\Delta n^0(x)}{D_e \tau_e} = 0 \quad (\text{A3.3})$$

and

$$\frac{\partial^2}{\partial x^2} \Delta n^1(x) - \frac{\Delta n^1(x)}{D_e \tau_e} (1 + 2\pi i f \tau_e) = 0 \quad (\text{A3.4})$$

Both of these equations have general solutions of the form

$$\Delta n(x) = A \exp(-x/L_e) + B \exp(x/L_e) \quad (\text{A3.5})$$

where A and B are constants and $L_e = \sqrt{D_e \tau_e}$. At $x = \infty$ we must have $\Delta n^0(\infty) = \Delta n^1(\infty) = 0$ and hence $B = 0$. The solutions to eqs (A3.3) and (A3.4) can therefore be written

$$\Delta n^0(x) = \Delta n^0(0) \exp(-x/L_e) \quad (\text{A3.6})$$

and

$$\Delta n^1(x) = \Delta n^1(0) \exp(-x/L_e^*) \quad (\text{A3.7})$$

where

$$(L_e^*)^2 = \frac{D_e \tau_e}{1 + 2\pi i f \tau_e} \quad (\text{A3.8})$$

We may define a frequency-dependent relative quantum efficiency $Q(f)$ by

$$Q(f) = P(f)/F(f) \quad (\text{A3.9})$$

Here $P(f)$ and $F(f)$ are the frequency-dependent parts of the photon generation rate within the material and the electron flow rate into the material respectively.

From eq. (A3.2),

$$F(f) = D_e \left[\frac{\partial}{\partial x} \Delta n^1(x) \right]_{x=0} = \frac{D_e \Delta n^1(0)}{L_e^*} \quad (\text{A3.10})$$

The recombination rate in a region of width δx is given by $\Delta n(x)\delta x/\tau_c$ (see eq. 2.38), and hence the total recombination rate throughout the material is given by

$$\frac{1}{\tau_c} \int_0^\infty \Delta n(x) dx$$

If we assume that each recombination gives rise to a photon then

$$P(f) = \frac{1}{\tau_c} \int_0^\infty \Delta n^1(x) dx$$

Hence, by substituting for $\Delta n^1(x)$ from eq. (A3.7),

$$P(f) = \frac{1}{\tau_c} \int_0^\infty \Delta n^1(0) \exp(-x/L_e^*) dx = \frac{L_e^* \Delta n^1(0)}{\tau_c} \quad (\text{A3.11})$$

From eqs (A3.9), (A3.10) and (A3.11) we then have

$$Q(f) = \frac{(L_e^*)^2 \Delta n^1(0)}{\tau_c D_e} = \frac{1}{1 + 2\pi i f \tau_c}$$

The modulus of this expression is $1/(1 + 4\pi^2 f^2 \tau_c^2)^{1/2}$ and hence we see that the light output $R(f)$ resulting when the current is modulated at a frequency f can be written

$$R(f) = \frac{R(0)}{(1 + 4\pi^2 f^2 \tau_c^2)^{1/2}}$$

Interactions between radiation and electronic energy levels with finite frequency linewidths

In section 5.2 we derived a relationship between the Einstein coefficients (eq. 5.1) by analyzing the interaction between a system with only two electronic energy levels and blackbody radiation. It was assumed that the energy levels were infinitesimally narrow and that only radiation at a 'single' frequency was involved. In the first part of this appendix we demonstrate that eq. (5.1) remains valid when the energy levels have a finite width.

We start by considering those transitions within the atomic system which give rise to emission frequencies between ν and $\nu + d\nu$ (Fig. A4.1). Suppose that the number of atoms per unit volume in the lower energy state which can absorb radiation with frequencies between ν and $\nu + d\nu$ is written $dN_1(\nu)$ whilst the number in the upper state which can give rise to stimulated emission over the same frequency range is written $dN_2(\nu)$. The number of absorption and stimulated emission transitions will then be proportional to the density of radiation which lies between ν and $\nu + d\nu$, that is to $\rho(\nu, T) d\nu$ (eq. 1.42). Thus we may write the number of upward transitions within this frequency range and per unit volume as $dN_1(\nu)\rho(\nu, T) d\nu B_{12}$. Similarly the corresponding number of stimulated emission transitions can be written $dN_2(\nu)\rho(\nu, T) d\nu B_{21}$. Spontaneous emission transitions will also contribute to the total number of downward transitions. However, we must remember that *all* the atoms in the excited state (i.e. N_2) can potentially cause photon emission between ν and $\nu + d\nu$ so that the probability that a given atom will do so will in fact be proportional to the factor $g(\nu) d\nu$, where $g(\nu)$ is the lineshape function. Thus we can write the spontaneous emission contribution as $N_2 g(\nu) d\nu A_{21}$.

Since the system is in equilibrium the upward and downward rates involving photon frequencies between ν and $\nu + d\nu$ must be equal and hence we have

$$dN_1(\nu)\rho(\nu, T) d\nu B_{12} = dN_2(\nu)\rho(\nu, T) d\nu B_{21} + N_2 g(\nu) d\nu A_{21} \quad (\text{A4.1})$$

We now integrate both sides over the width of the emission line. If we make the very reasonable assumption that neither the blackbody radiation density function nor the Einstein coefficients will vary appreciably over the narrow frequency range of the emission line, the result is

$$\rho(\nu_0, T) B_{12} \int_0^\infty dN_1(\nu) d\nu = \rho(\nu_0, T) B_{21} \int_0^\infty dN_2(\nu) d\nu + N_2 A_{21} \int_0^\infty g(\nu) d\nu \quad (\text{A4.2})$$

where ν_0 is the frequency corresponding to the centre of the emission line. Using the

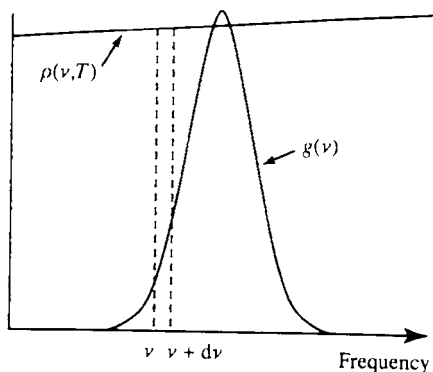


FIG. A4.1 An atomic transition between two energy levels with a lineshape function $g(\nu)$ is bathed in blackbody radiation which has a radiation density of $\rho(\nu, T)$.

relationships

$$\int_0^{\infty} dN_1(\nu) d\nu = N_1 \quad \int_0^{\infty} dN_2(\nu) d\nu = N_2 \quad \text{and} \quad \int_0^{\infty} g(\nu) d\nu = 1$$

we now obtain

$$\rho(\nu_0, T)B_{12}N_1 = \rho(\nu_0, T)B_{21}N_2 + A_{21}N_2$$

This expression will be valid whatever the value of ν_0 , and so we may replace it by a general frequency ν ; for notational convenience we also replace $\rho(\nu, T)$ by ρ_ν . Thus we have

$$N_1\rho_\nu B_{21} = N_2\rho_\nu B_{21} + N_2A_{21} \quad (\text{A4.3})$$

which is identical to eq. (5.1).

When dealing with the absorption of radiation caused by an atomic transition within an atom in section 5.3 it was assumed both that the radiation concerned was perfectly monochromatic and that the atomic energy levels were infinitesimally narrow. Here we investigate the consequences of allowing both the beam and the atomic transition to have finite frequency widths, although, in contrast to the situation considered in the first part of this appendix, we assume that the beam frequency bandwidth is much narrower than the transition linewidth (Fig. A4.2).

The function $g(\nu) d\nu$ gives the probability that a particular atom is capable of absorbing or emitting over the frequency range ν to $\nu + d\nu$, so the number of atoms per unit volume which are capable of absorbing radiation between these frequencies can be written as $N_1 g(\nu) d\nu$. In the presence of a radiation density of $\rho(\nu)$ the number of upward transitions per second per unit volume between these frequency limits can then be written as $N_1 g(\nu) d\nu B_{12} \rho(\nu)$. We now integrate over the frequency range of the incoming beam to obtain the total upward transition rate, R_{12} , due to absorption from the beam:

$$R_{12} = \int_0^{\infty} N_1 B_{12} g(\nu) \rho(\nu) d\nu \quad (\text{A4.4})$$

Since the beam linewidth is assumed to be small compared with the atomic transition

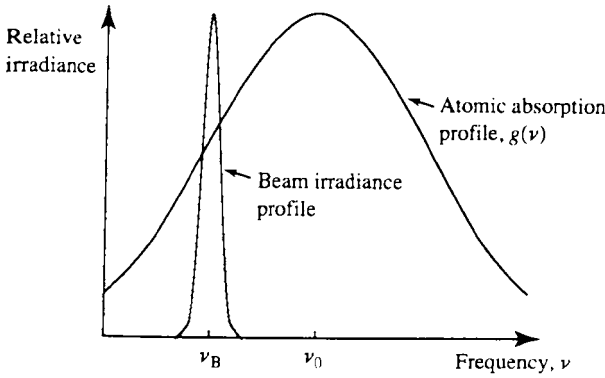


FIG. A4.2 A beam of radiation with a relatively narrow linewidth centred on ν_B passes through a collection of atoms where there is an atomic transition having a lineshape function $g(\nu)$ centred on ν_0 .

linewidth, the function $g(\nu)$ can be brought out of the integral and replaced by $g(\nu_B)$, where ν_B is the centre frequency of the beam. Thus we have

$$R_{12} = N_1 B_{12} g(\nu_B) \int_0^\infty \rho(\nu) d\nu \quad (\text{A4.5})$$

The integral on the right-hand side of eq. (A4.5) is simply the energy density of the entire beam, which we write as ρ_B , so that we have $R_{12} = N_1 B_{12} g(\nu_B) \rho_B$. Similarly we may write the downward transition rate due to stimulated emission as $R_{21} = N_2 B_{21} g(\nu_B) \rho_B$. Thus the net rate of loss of photons per unit volume, $-dN/dt$, from the beam as it travels through a volume element of medium of thickness Δx and unit cross-sectional area (Fig. 5.3) can be written

$$-\frac{dN}{dt} = N_1 \rho_B B_{12} g(\nu_B) - N_2 \rho_B B_{21} g(\nu_B) \quad (\text{A4.6})$$

or, substituting for B_{12} and B_{21} from eq. (5.7),

$$-\frac{dN}{dt} = \left(\frac{g_2}{g_1} N_1 - N_2 \right) \rho_B B_{21} g(\nu_B)$$

Since the beam can have any frequency, it is convenient simply to refer to it as having a frequency ν , so that we put $\rho_B = \rho_\nu$ and $g(\nu_B) = g(\nu)$, to give

$$-\frac{dN}{dt} = \left(\frac{g_2}{g_1} N_1 - N_2 \right) \rho_\nu B_{21} g(\nu) \quad (\text{A4.7})$$

Thus we obtain an expression similar to eq. (5.11) except that the right-hand side is multiplied by the function $g(\nu)$. It follows that eq. (5.13) should similarly be modified to read

$$\alpha(\nu) = \left(\frac{g_2}{g_1} N_1 - N_2 \right) \frac{B_{21} h \nu g(\nu)}{c} \quad (\text{A4.8})$$

Under conditions of population inversion the small signal gain coefficient can then be written as

$$k(\nu) = \left(N_2 - \frac{g_2}{g_1} N_1 \right) \frac{B_{21} h \nu g(\nu)}{c} \quad (\text{A4.9})$$

Optical bandwidths and pulse broadening

Let us suppose that at the input end of some optical system we have a signal where the optical power is modulated at a frequency f . That is, $P_{\text{in}} = P_0^i [1 - \cos(2\pi ft)]$. If we write the optical power output of the system as $P_{\text{out}} = P_0^o [1 - \cos(2\pi ft + \phi)]$, then the ratio of P_0^o to P_0^i gives the system response, which we denote by $H(f)$. We define the *optical bandwidth* of the system as the frequency range Δf_{opt} over which $|H(f)|$ exceeds one-half of its maximum value. If we allowed the optical output signal to fall onto a 'perfect' photodetector, then the current output, i_d , would follow the variations in $H(f)$. Electrical power from the detector, being proportional to i_d^2 , will follow $|H(f)|^2$. In electrical systems the *electrical bandwidth* is defined as the frequency range over which the *electrical power* exceeds one-half of its maximum value. We can therefore define an electrical bandwidth for our optical system, Δf_{el} , as the frequency range over which $|H(f)|$ exceeds $1/\sqrt{2}$ of its maximum value. Similar arguments apply when a voltage output is obtained.

We turn now to pulse broadening. Suppose we launch an optical pulse which, as far as the system response is concerned, is infinitesimally narrow. When the pulse emerges from the system, it will have broadened into a definite shape. This shape may be described by the *normalized impulse response* $h(t)$. The function $h(t)$ is normalized so that

$$\int_{-\infty}^{\infty} h(t) dt = 1$$

The functions $h(t)$ and $H(f)$ are in fact related by a Fourier transform (ref. A5.1), that is

$$H(f) = \int_{-\infty}^{\infty} h(t) \exp(-i2\pi ft) dt \quad (\text{A5.1})$$

A useful parameter for characterizing the spread of a pulse is the r.m.s. pulse width. This is given by σ where

$$\sigma^2 = \int_{-\infty}^{\infty} t^2 h(t) dt - \int_{-\infty}^{\infty} t h(t) dt \quad (\text{A5.2})$$

If $h(t)$ is symmetrical about a time t_0 , then by measuring time from t_0 we ensure that the second term on the right-hand side of eq. (A5.2) is zero.

σ is a useful quantity because it may be shown (ref. 8.21) that any pulse having an r.m.s. pulse width σ_1 that passes through a linear system whose impulse response has an r.m.s. width

σ_2 will emerge with an r.m.s. pulse width σ_3 where

$$\sigma_3^2 = \sigma_1^2 + \sigma_2^2 \quad (\text{A5.3})$$

By using eq. (A5.1) in conjunction with the definitions for Δf_{opt} and Δf_{el} , it is possible to determine expressions for the two latter quantities provided the function $h(t)$ is known. To give a relatively simple example of this type of calculation, we may take a rectangular normalized impulse response (Fig. A5.1a). That is, we have

$$h(t) = \frac{1}{\Delta T} \quad -\frac{\Delta T}{2} < t < \frac{\Delta T}{2}$$

and

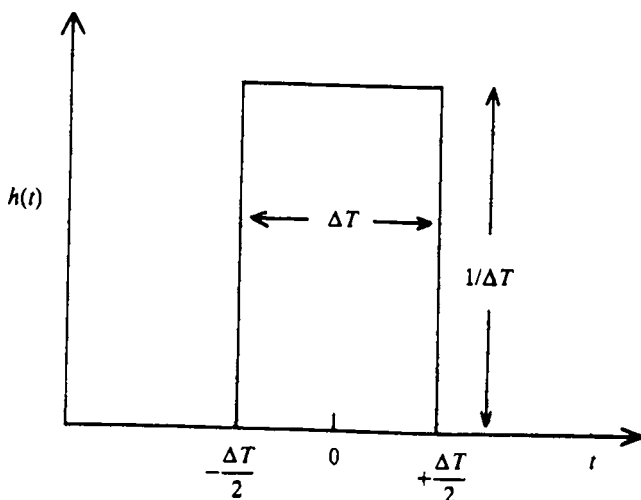
$$h(t) = 0 \quad t < -\frac{\Delta T}{2} \quad t > \frac{\Delta T}{2}$$

From eq. (A5.1)

$$\begin{aligned} H(f) &= \frac{1}{\Delta T} \int_{-\Delta T/2}^{\Delta T/2} \exp(-i2\pi ft) dt \\ &= \frac{i}{2\pi f \Delta T} \left[\exp\left(-\frac{i2\pi f \Delta T}{2}\right) - \exp\left(\frac{i2\pi f \Delta T}{2}\right) \right] \\ &= \frac{\sin(\pi f \Delta T)}{\pi f \Delta T} \end{aligned}$$

Thus $|H(f)|$ falls to one-half of its maximum value (i.e. 1 at $f=0$) when $\pi f \Delta T = 1.895$, that is $\Delta f_{\text{opt}} = 0.603/\Delta T$. Similarly, $\Delta f_{\text{el}} = 1.392/(\pi \Delta T)$ or $\Delta f_{\text{el}} = 0.443/\Delta T$.

(a)



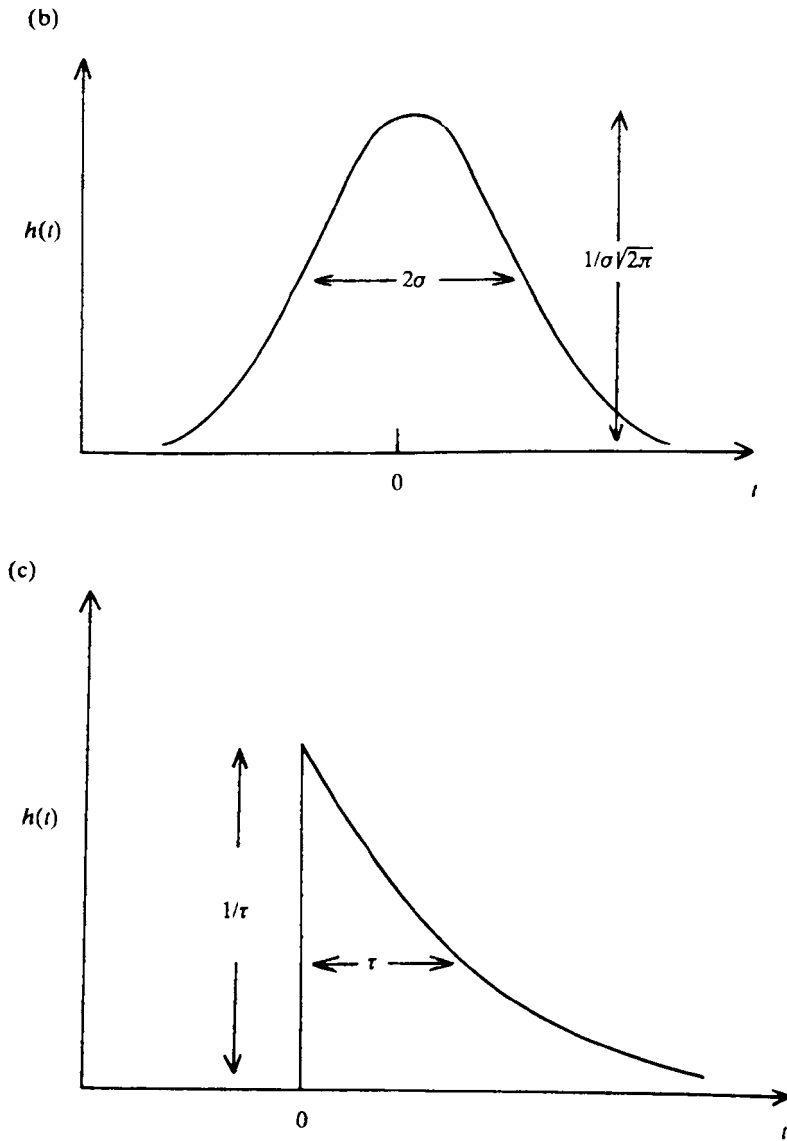


FIG. A5.1 Three 'theoretical' system response functions: (a) rectangular, (b) Gaussian and (c) exponential decay.

From eq. (A5.2) we also have

$$\sigma^2 = \frac{1}{\Delta T} \int_{-\Delta T/2}^{\Delta T/2} t^2 dt$$

Thus $\sigma^2 = \Delta T^2/12$ and hence $\sigma = \Delta T/2\sqrt{3}$, and we can therefore also write

$$\Delta f_{\text{opt}} = 0.174/\sigma \quad \text{and} \quad \Delta f_{\text{cl}} = 0.128/\sigma$$

When we are dealing with digital signals, the bandwidth is referred to in terms of B , the *bit rate*. We are then concerned with how close a succession of optical pulses can become before they can no longer be resolved (see Fig. A5.2). It may be shown (ref. A5.1) that for a wide range of pulse shapes B should not exceed $1/4\sigma$. Thus we can write for rectangular pulses that

$$\Delta f_{\text{opt}} \approx 0.7B \quad \text{and} \quad \Delta f_{\text{cl}} \approx 0.5B$$

Of course, all the above relationships between Δf_{opt} , Δf_{cl} , σ and B are only valid for rectangular pulses. For other pulse shapes, the relationships will be slightly different. A pulse shape that is often convenient to use is the Gaussian shape, for which

$$h(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-t^2/2\sigma^2)$$

This is illustrated in Fig. A5.1(b). We may proceed as for rectangular pulses, although the integrals involved are more difficult to evaluate, and we merely quote the results:

$$H(f) = \exp(-2\pi^2 f^2 \sigma^2)$$

and hence

$$\Delta f_{\text{opt}} = 0.187/\sigma$$

$$\Delta f_{\text{cl}} = 0.132/\sigma$$

Since we do not often know exactly what pulse shapes we are dealing with, it is convenient to use the following approximations:

$$\Delta f_{\text{opt}} \approx 2\Delta f_{\text{cl}}$$

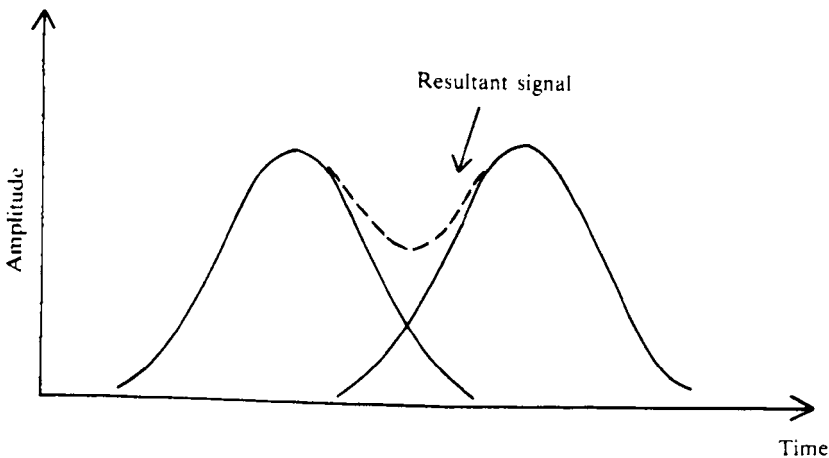


FIG. A5.2 Two overlapping pulses giving rise to a resultant signal that may be difficult to resolve into individual pulses if these are too close together.

$$B \approx 2\Delta f_{\text{el}}$$

$$B \approx \frac{1}{4\sigma}$$

$$\Delta f_{\text{el}} \approx \frac{1}{8\sigma}$$

Finally, another system response often encountered is that of exponential decay (Fig. A5.1c). In this case we may write

$$h(t) = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) \quad t > 0$$

$$h(t) = 0 \quad t < 0$$

From eq. (A5.1)

$$\begin{aligned} H(f) &= \frac{1}{\tau} \int_{-\infty}^{\infty} \exp\left(-\frac{t}{\tau}\right) \exp(-i2\pi ft) dt \\ &= \frac{1}{\tau} \int_{-\infty}^{\infty} \exp\left[-t\left(\frac{1}{\tau} + i2\pi f\right)\right] dt \\ &= \frac{1}{1 + i2\pi f\tau} \end{aligned}$$

whence

$$|H(f)| = \frac{1}{(1 + 4\pi^2 f^2 \tau^2)^{1/2}} \quad (\text{A5.4})$$

and

$$\Delta f_{\text{opt}} = 0.276/\tau$$

$$\Delta f_{\text{el}} = 0.159/\tau$$

REFERENCES

- A5.1 J. Gowar, *Optical Fiber Communication Systems*, Prentice Hall International, Hemel Hempstead, 1984, Section 2.4.2.

Physical constants and properties of some common semiconductors at room temperature (300 K)

Rest mass of electron	m	$= 9.110 \times 10^{-31} \text{ kg} = 0.000\,549 \text{ amu}$
Charge of electron	e	$= 1.602 \times 10^{-19} \text{ C}$
Electron charge/mass ratio	e/m	$= 1.759 \times 10^{11} \text{ C kg}^{-1}$
Avogadro's constant	N_A	$= 6.022 \times 10^{23} \text{ mol}^{-1}$
Planck's constant	h	$= 6.626 \times 10^{-34} \text{ J s}$
	$\hbar = h/2\pi$	$= 1.055 \times 10^{-34} \text{ J s}$
Boltzmann's constant	k	$= 1.381 \times 10^{-23} \text{ J K}^{-1}$
Speed of light (in vacuum)	c	$= 2.998 \times 10^8 \text{ m s}^{-1}$
Permittivity of a vacuum	ϵ_0	$= 8.854 \times 10^{-12} \text{ F m}^{-1}$
Permeability of a vacuum	μ_0	$= 4\pi \times 10^{-7} = 1.258 \times 10^{-6} \text{ H m}^{-1}$
Stefan-Boltzmann constant	σ	$= 5.670 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$

Property	Si	Ge	GaAs
Atomic (molecular) weight	28.09	72.60	144.6
Energy gap, E_g (eV)	1.12	0.67	1.43
Intrinsic carrier concentration, n_i (m^{-3})	1.5×10^{16}	2.4×10^{19}	1×10^{13}
Electron mobility, μ_e ($\text{m}^2 \text{ V}^{-1} \text{ s}^{-1}$)	0.135	0.39	0.85
Hole mobility, μ_h ($\text{m}^2 \text{ V}^{-1} \text{ s}^{-1}$)	0.048	0.19	0.045
Relative permittivity, ϵ_r	11.8	16.0	10.9 [†]
Electron effective mass [‡] , m_e^* ($\times m$)	0.12	0.26	0.068
Hole effective mass [‡] , m_h^* ($\times m$)	0.38	0.23	0.56
Recombination constant, B ($\text{m}^3 \text{ s}^{-1}$)	1.79×10^{-21}	5.25×10^{-20}	7.21×10^{-16}

[†] The range of values of relative permittivity of GaAs quoted in the literature is from 10.7 to 13.6. We have adopted the value 10.9.

[‡] Two different 'effective masses' are defined, namely the *density of states* and *conductivity* effective masses. The values given here are representative of those quoted for the conductivity effective mass.

Laser safety

Most practical lasers emit radiation that is potentially hazardous. The degree of hazard is related to the output characteristics of the laser, the way in which it is used, and the experience of the operator. The types of laser and the ways in which they are used are, as we have seen, many and varied. It would therefore be unreasonable to suggest that all laser applications require the same degree of control. Accordingly, systems of classification of lasers have been introduced, which are given below.

Any classification must be operable and consistent with our knowledge concerning the injuries that laser radiation can cause. The mechanism by which laser radiation causes damage is similar for all biological systems and may involve thermal, thermoacoustic and photochemical processes. The degree to which any of these mechanisms is responsible for damage depends on the parameters of the laser source, such as wavelength, pulse duration, image size and power, and energy density.

One of the principal characteristics of laser radiation is its beam collimation. This, together with a high energy content, can result in large amounts of energy being transmitted to biological tissues. The primary event causing damage is the absorption of the radiation by the biological system. Absorption occurs at an atomic or molecular level and is thus wavelength specific. Thus it is primarily the laser wavelength which determines which tissue is liable to be damaged. Absorption of radiation leads to a rise in temperature which may then disrupt molecular bonds and hence impair the function of the molecule. The rise in temperature is related to the length of exposure and the irradiance of the beam, and this leads to the notion of a radiant exposure, or dose, measured in joules per square metre. Exposure to *Q*-switched or mode-locked lasers can cause a very rapid rise in temperature of the tissue, so that liquid components may be converted to gas leading to a rupturing of cells. In general, the interrelationships between damage mechanisms and exposure are very complex (ref. A7.1), and we shall not pursue them here. The maximum permissible exposure (MPE) levels for which the human eye will not suffer adverse effects from exposure to laser radiation under certain conditions are given in ref. A7.2a, which gives guidelines and recommendations for the UK and rest of Europe.

Probably the most vulnerable part of the body as far as laser radiation is concerned is the eye. This is mainly because the eye lens will focus incident collimated laser radiation to a small point with a radius of the order of the wavelength, and with a corresponding high energy density. The hazards are wavelength dependent so that radiation in the ultraviolet and infrared regions, which is absorbed by the cornea, represents a danger to the cornea, whilst radiation

in the visible and near-infrared represents a retinal hazard. The increase in irradiance from the cornea to the retina is given approximately by the ratio of the pupil area to that of the image formed on the retina. Typically, the pupil may expand to a diameter of 5 mm with a corresponding retinal image of 10–20 μm diameter. Thus the irradiance increases by a factor of between 2×10^5 and 5×10^5 . A laser beam of 50 W m^{-2} incident on the cornea then represents an irradiance of about 25 MW m^{-2} on the retina. About 5% of the radiation falling on the retina is absorbed in the pigments within the rods and cones, and this may be sufficient to cause a burn and a loss of vision.

In general terms, skin can tolerate a great deal more exposure than can the eye, although blisters, ulceration and scarring of the skin may occur for high levels of irradiation. Again, the hazards are wavelength dependent, being more pronounced for ultraviolet radiation. Any organization which uses lasers would be advised to draw up a safety code of practice, which should be based on an accepted classification of lasers. The classification accepted in the UK and Europe is described below. There are slight differences in the classification used in the United States. These classifications and related information on laser safety are given in considerable detail in the various publications cited in ref. A7.2.

A7.1 Laser classification based on BS 4803

Class 1

Output power is so low as to be inherently safe.

Class 2

Such lasers operate in the visible part of the spectrum (400 nm to 700 nm) and their output power is limited to 1 mW for continuous wave (CW) operation. Such lasers are not inherently safe, but some eye protection is afforded by the natural aversion response of the eye, including the blink reflex. Hazards can be controlled by relatively simple procedures.

Class 3A

These lasers operate in the visible part of the spectrum (400 nm to 700 nm) and their output is limited to 5 mW for CW operation. Some protection is still provided by the aversion responses. Direct intrabeam viewing with optical aids may be hazardous.

Class 3B

These lasers operate in any part of the electromagnetic spectrum between wavelengths of 200 nm and 1 mm. Their output power is limited to 500 mW for CW operation. Direct beam viewing could be hazardous and must be avoided. Likewise specular reflections may be hazardous but diffuse reflections will not generally be so. Under no circumstances should the beam be viewed with optical aids. More detailed control measures are necessary.

Class 4

These lasers also operate within the wavelength range 200 nm to 1 mm and their output power exceeds 500 mW. Not only is viewing of the direct beam and specular reflections hazardous but, in some cases, viewing of diffuse reflections may also be hazardous to the eye. In addition, there is also a risk of skin burns from the direct beam and from first-order specular reflections. The beam from such lasers is also capable of igniting flammable material so that care must be taken to minimize the risk of fire. The use of Class 4 lasers requires extreme caution for safety both of the user and of other persons who may be present. If possible the system should be totally enclosed.

A7.2

Safety note

It is stressed that the above classification is intended as a guide only and that laser users should consult the publications given in ref. A7.2. Particular attention is drawn to the American 'Standards for the safe use of lasers' Number Z136 issued by the National Standards Institute (ref. A7.2b).

Each laser device should carry an appropriate label and notice of warning (ref. A7.2a, Part 2, p. 3).

The safe use of lasers often involves the provision of safety interlocks and warning lights on access doors to rooms where lasers are being used, together with beam stops and enclosures. Materials which act as diffuse reflectors should be used wherever possible. Protective eyewear should also be available for use with Class 3B and Class 4 lasers. Bearing in mind that the protective medium will not afford protection over the whole spectral range, care must be taken to ensure that the eye protection has the correct spectral response to match the laser being used.

Finally, it should be remembered that most lasers are high voltage devices (several kilovolts) often capable of delivering large currents and therefore care must also be taken in this respect.

REFERENCES

- A7.1 A. F. McKinlay and F. Harlon, 'Biological bases of maximum permissible exposure levels of laser standards. Part 1: Damage mechanisms', *J. Soc. Radiol. Prot.*, **4** (1), 1984.
- A7.2 (a) BS EN 60825-1:1994 'Safety of Laser Products'. (A software package, Laser Safe PC, attempts to help interpret these regulative procedures, by GL Services, available from Lasernet Ltd, Brockenhurst, Hants, UK, and incorporating BS EN 60825-1 and BS EN 60825-2, which includes recommendations for fiber optic work.)
- (b) American National Standards Institute, 'Standards for the safe use of lasers', *ANSI Z136*, 1980.
- (c) International Electrotechnical Commission, IEC Standard 'Radiation safety of laser products, equipment classification, requirements and users' guide', Publication 825, 1984.

- (d) CVCP of Universities in UK: Safety in Universities, Notes for Guidance, Part 2:1 Lasers (3rd edn), 1992.
- (e) R. J. Rockwell, 'Analysing laser hazards', *Lasers Appl.*, May, 97, 1986.
- (f) J. Kaufman and R. Tucker, 'Laser absorbers for eye protection', *Lasers Appl.*, Oct., 69, 1985.
- (g) D. C. Winban, 'Dispelling myths about laser eyewear', *Lasers Appl.*, Mar., 73, 1986.
- (h) R. J. Rockwell, 'Controlling laser hazards', *Lasers Appl.*, Sept., 93, 1986.
- (i) M. L. Wolbarsht and D. H. Sliney, 'Laser safety standards move on', *Lasers Appl.*, Apr., 97, 1987.

Index

- aberrations (thin lens), 25
- absorption
 - coefficient, 173
 - effects on detector performance, 318
 - variation in semiconductors, 316
 - loss in fibers, 395–7
 - of radiation, 173–5
- acceptors, 55
 - acting as traps, 130
- acoustic phonon, 501
- acousto-optic
 - beam deflection, 480, 484
 - effect, 112–16
 - modulation, 116
- activators, 129
- active displays, 129
- active mode locking, 252–4
- active phase tracking, 504–6
- active region, 205
- alexandrite laser, 202–3
- amplifier front end design, 454–6
 - high impedance, 454
 - low impedance, 454
 - transimpedance, 454–6
- amplitude shift keying (ASK), 435
- analog modulation, 429–32
- anamorphic lens, 445
- anisotropic material, 527
- anisotype junction, 80
- antireflection coating
 - fiber bundles, 521
 - Schottky photodiodes, 335
 - silicon photodiodes, 329
- anti-Stokes shift, 501
- argon ion laser, 226–8
- array
 - CCD 2D, 347
 - detector, 344–8
 - laser, 219–20
 - phased, 220
 - scanning, 164–5, 344
 - VCSEL, 221–2
- articulated arm, 280
- aspect ratio, 285
- atmospheric
 - turbulence, 436
- window, 436
- attenuation in optical fibers, 388, 395
- measurement, 407–8
- avalanche
 - breakdown, 79
 - photodiode, 337–40
 - amplification process, 337–8
 - in communication systems, 438, 462
 - in OTDR, 412
 - noise, 340–1
 - reach-through, 339
 - separate absorption and multiplication (SAM), 340
 - structure, 338–9
- axial modes, 190–3
- backscatter (fibers), 411
- balanced receiver configuration, 469, 486
- band bending, NEA surfaces, 304
- bandgap *see* energy gap
- bandgap wavelength, 86, 117, 314
- bandwidth
 - electrical, 541
 - optical, 541
- beam
 - coherence, 260–4
 - collimation, 259
 - divergence, 258–60
 - table, 259
 - modulation telemetry, 269–71
- beam splitter (fiber), 510, 518
- beat length, 418, 513
- beta-barium borate, 237
- birefringence
 - fibers, 418, 512–13
 - induced, 532–4
 - natural, 527–32
- birefringent filter, 204
- bistable optical devices, 482–4
 - logic gates, 482
 - optical computer, 482
 - R-SEED devices, 482
 - S-SEED devices, 483
 - single rail logic, 484
- bit, 433
 - error rate, 434–5, 452, 464
- blackbody radiation, 28–30
 - Planck's law, 30
 - Wien's law, 30
- Bohr model, 31–2
- boiling point, 281
- bolometer, 298–300, 348
- Boltzmann statistics, 175–7
- Bragg
 - acousto-optic grating, 114–16
 - fiber grating, 420–1, 467, 484, 498–9
 - as sensor, 498–9
 - hydrogen loading, 420–1
 - notch filter, 467
 - phase grating, 421
 - strain sensitivity, 498
 - grating reflector, 480, 486
- Brewster angle, 8
- Brewster law, 8
- brightness
 - displays, 157–8, 165
 - laser radiation, 264–5
 - LEDs, 157–8
- Brillouin scattering, 468, 501
- Brillouin zone, 45
- Burrus emitter, 444
- cadmium selenide (CdSe)
 - photodetectors, 319
- cadmium sulfide (CdS) photodetectors, 319
- cadmium telluride, 107, 354
- capacitance, junction, 75–8
 - effect on LED response time, 152
 - effect on photodiode response time, 333
- capacitor, MOS, 344
- carbon dioxide laser, 229–32
 - gas dynamic laser, 232
 - gas flow laser, 232
 - materials processing, 279, 281–3
 - rotational–vibrational energy levels, 229–30
 - sealed tube laser, 231
 - surgery, 418
 - TEA laser, 232

- carrier
 - concentrations, 57–62
 - confinement, 81
 - lifetime, 152–3
- cathode ray tube (CRT), 134–7
 - electron gun, 135
 - faceplate, 136
 - raster scanning, 135–6
 - screen, 135
 - shadowmask, 136
 - structure, 135
- cathodoluminescence, 133–7
 - definition, 129
 - energy transfer mechanism, 134
 - use in CRT screen, 135
- cavity
 - lifetime, 256
 - optical, 179–80
- chalcogenide glasses, 418
- characteristic luminescence, 130
- charge
 - carriers, 47
 - coupled device (CCD), 344–8
 - arrays, 351
 - polysilicon gate, 347–8
 - 2D array, 347
 - storage capacitance, 77
 - transfer efficiency, 346
- chemical vapour deposition
 - axial (VAD), 414
 - modified (MCVD), 413
 - outside (OVD), 414
- chirped diffraction grating, 480
- cholesteric liquid crystals, 159
- chromatic aberration, 25
- cladding
 - elliptical, 420
 - mode, 377, 406
 - pumping, 202
 - waveguides, 359, 365, 474
- cleaved coupled cavity, 249
- co-activators, 130
- coded mark inversion, 435
- coherence, 260–4
 - length, 262, 510
 - in frequency doubling, 121
 - of laser light, 260–4
 - spatial, 262
 - temporal, 262
 - time, 263
- coherent fiber bundles, 519–21
- coherent optical fiber communication
 - systems, 464, 468–70
 - balanced receiver configuration, 469
 - polarization diversity, 469
 - polarization scrambling, 469
- common emitter gain, phototransistor, 342
- compact disc, 487
- compound lens, 26
- conduction band, 43, 47–8
- cone (eye), 548
- contact potential, 69
- conservation of wavevector in
 - interband transitions, 142–3
- copper indium diselenide, 354
- core
 - elliptical fibers, 419, 513
 - waveguides, 359, 365, 474
- cornea, 547
 - sculpting, 286
- corner cube reflector, 269
- coupled mode theory, 405, 460, 478
- couplers
 - multifiber, 403–5
 - single mode, 405–6
- cross-talk (fibers), 519–20
- crystal
 - class, 529
 - field, 197, 200
- Curie temperature, 107–8, 301
- cut-off frequency in photodetectors, 294, 301
- cut-off wavelength (optical fibers), 409–10
- D^* , 294–5, 344
- dark-line defects, 217
- de Broglie equation, 37
- decision level, 434
- decision time, 434, 464
- degeneracy, 172, 240
- degradation (lasers), 217–18
- density of states function, 57
 - quantum well, 85
- depletion layer
 - capacitance, 75–7
 - width, 328
- depletion region, 69, 81, 325, 333
- depressed cladding fibers, 392
- difference frequency generation, 125
- diffraction, 480
 - grating, 467
 - grating (chirped), 480
 - limited, 26
- diffusion
 - capacitance, 77, 152
 - coefficient, 65
 - effect on LED response, 152–3
 - length, 65, 205
 - of carriers, 64–6
 - in relation to photodiode response, 334
 - potential, 69
- digital modulation, 432–6
 - fundamental limitations on signal size, 451–4
 - noise, 433, 452
- dipoles (electric), 301, 527
- direct bandgap, 46, 142–3, 205
- direct detection, 429, 452, 470
- directionality of laser radiation, 258–60
- director (liquid crystal), 159, 162
- dispersion
 - compensation, 462
 - flattened fibers, 392, 462
 - intermodal, 378–82, 383–4, 457
 - material, 389–91
 - measurement, 408–9
 - profile, 389, 391
 - shifted, 462
 - waveguide, 389, 391
- displays
 - active, 129
 - brightness, 157–8
 - passive, 129, 158
 - seven bar segment, 163
 - 7×5 matrix, 163
- distance measurement, 267–74
 - beam modulation telemetry, 269–70
 - interferometric methods, 267–8
 - pulse echo technique, 271
- distributed feedback laser, 221, 469
- document scanner, 518
- donors, 53–4
 - acting as traps, 130
- doped insulator laser, 196–204
 - diode pumped, 200–2
- doped semiconductor photodetector, 321
- doping, 53–6
- Doppler
 - broadening, 184
 - effect, 116, 246
- double refraction *see* birefringence
- drift (charge carriers), 50, 69
- dye (saturable), 253–4
- dye lasers, 233–6
- effective density of states, 60–1
- effective mass, 50, 56
- eigenfunctions, 41
- eigenvalues, 41, 42
- Einstein, 2
 - coefficient, 171, 188
 - relations, 172
 - lasers, 171–2
 - semiconductors, 65
- electric
 - dipoles, 301, 527
 - displacement, 527
 - permittivity tensor, 528
 - polarization, 119–20, 527–8
 - susceptibility, 527
 - tensor, 528
- electrical
 - bandwidth, 541
 - conductivity, 48–51
- electroluminescence, 129
 - a.c. powder device, 138–41
 - a.c. thin film device, 138, 141
 - classical, 129
 - d.c. powder device, 139–40
 - d.c. thin film device, 138
 - definition, 129

- injection, 138, 141–2
 - emission mechanisms, 139–40
- electromagnetic spectrum, 2
- electro-optic
 - coefficients, 96–7
 - effect, 96–107
 - modulator, 477–8
 - longitudinal, 103
 - transverse, 103–4, 476–7
- electron
 - accumulation, 80
 - affinity, 62, 80, 82, 304
 - beam range, 134
 - beam pumped lasers, 233
 - gun, 135
 - electron–hole pairs, 47
- emission (of radiation)
 - spontaneous, 170–1
 - stimulated, 170–1
- endoscope, 286, 519
- energy
 - bands, 42–8
 - gap, 43, 84
- energy-level systems (in lasers)
 - four level, 177–8, 187, 198, 203
 - three level, 178, 458
 - two level, 177, 200
- evanescent field, 364
- excess carriers, 63–4
- excess loss (fiber couplers), 403
- excess noise factor, 340
- excimer laser, 233
 - materials processing, 279, 282
- ophthalmology, 286
- excitons, 56–7, 117
 - bound, 145
 - in cathodoluminescence, 134
 - in LEDs, 145
- exponential decay, 294, 545
- external quantum efficiency, 151
- extra mural absorption (e.m.a.), 521
- extraordinary ray, 93–4, 123
- extraordinary refractive index, 94
- extrinsic semiconductors, 53–6
- eye diagram, 463–4
- Fabry–Perot
 - cavity, 125
 - interferometer, 18–19
 - resonator, 179, 192, 244
- Faraday
 - effect, 110–12
 - rotation, 513
- Fermi–Dirac function, 58
- Fermi energy level, 58, 62, 69, 80, 81, 83
- fiber
 - acceptance angle, 376, 441
 - bundles, 516–21
 - arrays, 519
 - beam splitters, 518
 - coherent, 286, 519–21
 - core packing fraction, 516
 - cross-talk, 519
 - endoscope, 286, 519
 - fiberscope, 519
 - in medicine, 286
 - matrix display, 518
 - mixer, 518
 - rigid, 519–21
 - Y-guide, 518
 - cables, 421–3
 - couplers, 412–16
 - dispersion measurement, 408–9
 - jointing, 397–402
 - losses, 393–7
 - absorption, 397
 - bending, 393–5
 - in plastic fiber, 416
 - jointing, 397–402
 - microbending, 394
 - mid-infrared fibers, 417
 - scattering, 359
 - manufacture, 413–16
 - fiberscope, 519
 - filaments in lasers, 211
 - flicker noise, 319
 - fluorescence, 129
 - fluorescent dyes, 234
 - fluorescent lamp, 133
 - fluoroalkyl methacrylate, 415
 - Fluoroptic sensor, 496
 - flux budget calculation, 462
 - focal length, 25
 - focal point, 24
 - focusing of laser radiation, 265–7
 - foot-Lambert, 158
 - Photonic sensor, 493
 - Fourier transform, 541
 - frame transfer, 346
 - Fraunhofer diffraction, 21–2
 - free electron laser, 238–9
 - free space communications, 436–8
 - frequency
 - doubling *see* second-harmonic generation
 - shift keying (FSK), 436, 468
 - stabilization in lasers, 245–9
 - Fresnel
 - diffraction, 21
 - loss at fiber joints, 398
 - reflection, 411, 521
 - Fresnel–Kirchhoff formula, 21
 - Fresnel's equations, 361
 - fused biconic taper coupler, 405
 - use as wavelength mixer, 460–1
 - fused plate, 520–1
 - fusion (laser induced), 286–8
 - fusion splicing, 397
 - Gabor, 271
 - gain
 - bandwidth product (GBP), 505
 - coefficient, 175
 - curve, 192, 245
 - saturation, 182, 468
 - gallium aluminum arsenide (GaAlAs)
 - heterojunction, 80
 - IO waveguides, 475
 - lasers, 208, 211–15, 482
 - LEDs, 148
 - quantum wells, 83
 - VCSELs, 222
 - gallium arsenide (GaAs)
 - fiber sensors, 295
 - heterojunction, 80
 - induced birefringence, 532–4
 - intensified photodiode, 344
 - IO devices, 486
 - IO waveguides, 475
 - lasers, 205–7, 211
 - LEDs, 147
 - metal semiconductor metal photodiode, 337, 484
 - photocathodes, 313
 - Pockels modulators, 107
 - quantum wells, 83
 - Schottky photodiode, 336
 - solar cell, 354
 - VCSELs, 222
 - gallium arsenide phosphide (GaAsP)
 - intensified photodiode, 344
 - lasers, 215, 216
 - LEDs, 148
 - Schottky photodiode, 336
 - gallium indium arsenide (GaInAs)
 - APD, 340
 - IO devices, 486
 - photoconductor, 320
 - photodiodes, 331–2
 - gallium indium arsenide phosphide (GaInAsP)
 - APD, 457
 - IO devices, 486
 - lasers, 216–17
 - photodiodes, 331
 - gallium nitride, 222
 - gallium phosphide (GaP), 147, 336
 - gas lasers, 223–33
 - argon ion, 226–7
 - CO₂, 229–32
 - excimer, 233
 - He–Ne, 223–6
 - metal vapour, 228
 - nitrogen, 233
 - gate (MOS device), 344
 - Gaussian
 - frequency distribution, 185
 - mode, 193–5, 387, 446
 - pulse, 408, 544
 - geodesic lens, 484
 - generation–recombination noise, 318
 - germanium
 - APD, 340, 456
 - photodiode, 311
 - glass fiber, 516

- amplifier, 202
- laser, 201–2
- glaucoma, 286
- Golay cell, 300
- graded index fibers, 382–5
 - intermodal dispersion, 383–5
 - light coupling from LED, 443
 - modes, 384–5
 - ray paths, 384–5
 - refractive index profile, 382
- GRIN lens, 26–8, 444
- group
 - index, 389
 - transit time, 409
 - velocity, 4–5, 389
- guard ring (APD), 338
- gyro (fiber), 508–11
- half-wave voltage, 100–3
- heat-affected zone, 283
- He–Ne laser, 223–6
- Heisenberg's uncertainty principle, 38, 144
- helical ray, 385
- heterodyne detection, 429–30, 468
 - balanced receiver, 486
- heterojunction, 79–81, 331
 - lasers, 211–18, 348
 - stripe geometry, 211–14, 217
- high birefringence fiber, 419–20
- hole burning, 245–6
- holes, 47
- hologram, 271
- holographic interferometry, 274–6
- holographic optical memories, 276–8
- holography, 271–4
- homeotropic ordering, 159
- homodyne detection, 431, 468
- homogeneous broadening, 185, 245
- homogeneous ordering, 159
- homojunctions, 67–70, 331
 - lasers, 204–7
- Huygens' construction, 94
- Huygens' principle, 19
- hydrogen loading, 421
- ideality factor, 79
- idler wave, 237
- image
 - conduit, 520
 - contrast, 521
 - intensifiers, 312–14, 521
 - first-generation types, 312–13
 - second-generation types, 313
 - third-generation types, 313–14
- impurity centre recombination, 144–5
- index
 - ellipsoid, 93, 529–30
 - matching, 122–3, 125
- indirect bandgap, 142–3
- indium antimonide (InSb)
 - photoconductor, 321
- indium gallium arsenide phosphide (InGaAsP) *see* gallium indium arsenide phosphide
- indium gallium nitride (InGaN), 222
- induced birefringence, 532–4
- inhomogeneous broadening, 185–6, 245–6
- injection luminescence, 141–2
- insertion loss, 402, 404
- integrated optics, 472–87
 - beam deflectors, 480
 - bistable devices, 482–4
 - couplers, 478–80
 - detectors, 481, 484–6
 - emitters, 481–2
 - filters, 480
 - modulators, 478–80
 - phase shifter, 476–8
 - quantum well modulator, 118
 - spectrum analyzer, 484
 - switch, 477–80
 - waveguides, 472–5
 - field distributions, 474–5
 - losses, 475
 - waveguide splitter, 475–6
- intensified photodiode, 343–4
- interband transition, 142–4
- interference, 14–19
 - fringes, 13, 274–5
 - multiple beam, 19
 - thin film, 16–17
 - two beam, 14–17
- interferometer
 - Mach–Zehnder, 503–4
 - Michelson, 268–9
- interferometry (holographic), 274–6
- interline transfer, 346
- intermodal
 - dispersion, 377–82, 383–5, 457
 - frequency separation, 190–1, 244
- internal quantum efficiency, 151, 210
- intrinsic semiconductors, 51–3
- ion implantation, 83
- ion lasers, 226–8
- isoelectronic trap, 148
- isotropic material, 527
- isotype junction, 80
- Johnson noise, 311, 448–50, 454, 456
- junction
 - anisotype, 80
 - capacitance, 77
 - geometry, 75
 - heterojunction, 79–81, 331
 - homojunction, 67–70, 331
 - isotype, 80
 - metal–semiconductor, 81–3, 334
 - metal–semiconductor–metal, 336, 484
 - ohmic, 80, 83
 - p–i–n, 329–30
 - p–n, 67–79
 - rectifying, 80
- kerf, 283
- Kerr
 - constant, 97
 - table, 107
 - effect, 96
 - magnetic, 112
 - optical, 108, 470
 - modulators, 107–8
- keyhole (laser welding), 282
- kinks (laser output), 211, 447
- krypton fluoride excimer laser, 233
- Kynar, 422
- Lamb dip, 247
- Lambertian emission, 157
 - coupling to fibers, 442
 - display brightness, 157
 - source, 410, 516
- laser
 - arrays, 219–20
 - beam delivery, 279–80, 418
 - cutting, 278, 283
 - heat-affected zone, 283
 - kerf, 283
 - drilling, 278, 284
 - aspect ratio, 285
 - excimer laser, 285
 - ruby laser, 284
 - taper, 285
 - induced fusion, 286–8
 - light (properties of), 258–67
 - brightness, 264–5
 - coherence, 260–4
 - directionality, 258–60
 - focusing properties, 265–7
 - tunability, 236, 267
 - losses, 181
 - marking, 278, 284
 - materials processing, 279
 - medical applications, 285–6
 - mode
 - frequency separation, 190–1
 - volume, 179–80, 193–4, 207
 - modes
 - axial, 190–3
 - transverse, 193–5
 - safety, 547–50
 - scribing, 284
 - vertical cavity surface emitting (VCSEL), 220–2
 - welding, 278, 281–2
 - shielding gas, 281
 - keyhole formation, 282
- lasers
 - alexandrite, 202–3
 - cadmium ion, 228
 - diode pumped, 200–1
 - doped insulator, 196–204
 - free electron, 238–9

- gas, 223–33
- glass fiber, 201–2
- krypton fluoride, 122
- liquid dye, 233–6
- metal vapour ion, 228–9
- Nd–glass, 199
- Nd–YAG, 196–9, 438
- parametric, 236–8
- quantum well, 218–19
- ruby, 199–200
- semiconductors, 204–22, 248–9
 - see also* semiconductor lasers
- short wavelength, 222
- tunable, 202, 238
- vibronic, 202–4
- xenon fluoride, 233
- latent heats, 281
- lattice matching, 222, 331
- lead sulfide (PbS) photoconductor, 319–20
- lead zirconate, 301
- leaky mode, 374, 406, 410
- lens
 - anamorphic, 445
 - compound, 26
 - geodesic, 484
 - GRIN, 26–8, 444
 - sphere, 26, 444, 445
 - thin spherical, 24–6
- lidar, 271
- lifetime, 170
- light-emitting diode (LED), 141–55
 - communication systems, 428, 438, 441, 461, 463
 - construction, 149–52
 - drive circuits, 153–5
 - edge emitting, 81, 444
 - emission linewidths, 146
 - impurity centre recombination, 144–5
 - injection luminescence, 141–6
 - materials for, 147–9
 - quantum well, 86
 - recombination processes in, 142–6
 - response times of, 152–3
 - surface emitting, 81
 - superluminescent, 222–3, 510–11
- line coding, 435
- linearly polarized light, 7–11
- linearly polarized mode, 374, 385
- lineshape function, 183, 537
- linewidth, 260
- liquid crystal
 - displays, 158–63
 - cell construction, 160–3
 - effect of applied voltage, 161
 - matrix, 487
 - supertwist type, 163
 - twisted nematic type, 160–3
- light valve, 348–51
 - accumulation phase, 351
 - depletion phase, 350
 - hybrid field effect mode, 349
 - large screen projection, 351
 - phase epitaxy, 475
 - pitch, 159
 - state, 159
- lithium niobate (LiNbO₃)
 - holography, 278
 - IO devices, 475, 477–8, 481
 - parametric laser, 237
 - parametric oscillation, 124
 - Pockels cell, 107, 257
- lithium tantalate, 107
- local area network (LAN), 465–6
 - components, 466
 - topologies, 465
- logic gate, 482
- Lorentzian lineshape, 185
- loss coefficient (lasers), 181–2, 209
- luminescence, 129–31
 - characteristic, 130, 134, 137
 - injection, 141
 - non-characteristic, 130, 134, 137
 - lifetime, 130
- Mach–Zehnder interferometer, 503–5
- machine vision, 281
- magnetic radiation, 239
- magneto-optic devices, 110–12
- magnetostrictive materials, 508
- magnification
 - lens, 25
 - mirror, 24
- Maiman, 195
- majority charge carrier, 53
- material dispersion, 389–91, 462
 - in SiO₂, 390
- materials processing (laser), 281–5
- matrix displays, 163–6
 - brightness, 165
 - fiber bundles, 518
 - ideal element characteristic, 165
 - scanning, 164
 - suitability of electroluminescent devices, 141
 - use of liquid crystals in, 165–6
 - wiring methods, 164
- Maxwell
 - electromagnetic theory, 529
 - equations, 360
- medical applications (lasers), 285–6
 - cancer, 286
 - corneal sculpting, 286
 - dentistry, 286
 - glaucoma, 286
 - hypermetropia, 286
 - myopia, 286
 - ophthalmology, 286
 - optical fibers, 286
 - photodynamic therapy, 286
 - photorefractive keratectomy, 286
 - plaque removal, 286
 - surgery, 285–6, 418
- mercury cadmium telluride (HgCdTe), 321, 323
- meridional ray, 374, 377, 384, 385
- MESFET amplifier, 486
- metal oxide semiconductor (MOS) capacitor, 344
- metal–semiconductor junction, 81–3
- metal–semiconductor–metal (MSM) photodiode, 336–7
- metal vapour ion laser, 228–9
- metastable state, 129, 177
- metrology, 267–71
- Michelson interferometer, 267–8
- microbends
 - effect on fiber losses, 394
 - effect on mode dispersion, 381
 - in cables, 422
 - sensors, 497
- microbolometer, 348
- microchannel plate, 313
- microfilter (CRT), 137
- mid-infrared fibers, 417–18
 - chalcogenide glasses, 418
 - ZBLAN, 417
- Mie scattering, 436
- minority charge carrier, 53
- mirror (spherical), 23–4
 - multilayer, 221
 - bandpass, 467
 - phase conjugate, 280
- mixer (fiber), 518
- mixing rod, 403, 466, 512
- mobility, 49
- mode
 - coupling, 381, 393
 - cut-off, 368, 378
 - wavelength, 387
 - dispersion, 378–80, 383–5
 - field diameter, 387
 - locking, 250–4
 - scrambler, 406
 - stripper, 406
 - volume, 180, 193–5, 207, 208
- modes
 - fibers, 374–5, 385
 - cladding, 377, 406
 - electric field distribution, 375
 - leaky, 374, 406, 410
 - linearly polarized, 374, 385
 - sensors, 497
 - single mode, 385–7, 389, 446, 457
 - laser, 190–5
 - single mode operation, 244–5
 - planar waveguides, 368–73
- modified chemical vapour deposition (MCVD), 413
- modulation, 428–36
 - analog, 425–32
 - digital, 432–6
 - frequency, 431, 436, 438
 - lasers, 447

- LEDs, 154, 445
- pulse, 429
- molecular lasers, 229–33
 - CO₂, 229–32, 438
 - excimer, 233
 - gas dynamic, 232
 - nitrogen, 233
- monochromatic aberration, 25
- monofibers, 516
- multilayer mirror, 221
- multiple quantum well, 482
 - detector, 323–4
- multiplication noise, 311
- mutual coherence function, 261
- myopia, 286
- negative
 - bandgap, 321
 - crystals, 122, 532
 - electron affinity, 303–5
- nematic liquid crystals, 159
- neodymium lasers,
 - diode pumped, 200–1
 - glass, 199
 - materials processing, 282
 - YAG, 196–9, 438
 - materials processing, 279, 282
 - Q-switched, 237
- nit, 157
- nitrogen laser, 233
- noise
 - APD receiver, 450–1
 - avalanche, 340–1
 - equivalent power (NEP), 294
 - excess noise factor (APD), 340–1, 450
 - flicker, 319
 - generation–recombination, 318–19
 - in digital systems, 451–4
 - Johnson, 311, 448, 456
 - multiplication, 311
 - photoconductors, 318–19
 - photomultiplier, 310–12
 - p–i–n receiver, 448–50
 - shot, 310
 - white, 294
- non-characteristic luminescence, 130, 137
- non-linear optics, 119–24
 - frequency doubling *see* second-harmonic generation
 - paramagnetic amplification, 124–6
- non-return-to-zero (NRZ), 432
- normalization condition, 41
- normalized film thickness (V), 369
- normalized impulse response, 541
- numeric displays, 163–6
- numerical aperture (NA), 376, 516
- object wave (holography), 273
- ohmic contacts, 83, 328, 354
- ohmic junction, 80, 83
- Ohm's law, 49
- optic axis, 93, 95, 238, 529
- optical
 - activity, 96–7
 - bandwidth, 541
 - cavity, 179–80, 237
 - computer, 482
 - confinement, 214
 - feedback, 179–81
 - fiber amplifiers, 457–61, 464
 - erbium in silica, 459, 469
 - gain profile, 459
 - praesodymium in ZBLAN, 461
 - pumping arrangements, 458–9
 - use in optical fiber
 - communication systems, 464
 - fiber communication systems
 - advantages, 439
 - choice of fibers, 457
 - coherent systems, 464, 468–70
 - components, 461–5
 - detectors, 448, 456–7
 - direct detection, 470
 - dispersion compensation, 462
 - emitter design, 438, 440–7
 - eye diagram, 463–4
 - flux budget, 462
 - limitations on signal size, 451–4
 - solitons, 471
 - transatlantic links, 464–5
 - wavelength division multiplexing, 466–8
 - wavelength used, 440
 - fiber sensors, 492–515
 - Bragg grating, 498–9
 - Brillouin scattering, 501
 - displacement, 492–3
 - distributed, 500–3
 - extrinsic, 492–7
 - Fluoroptic, 496
 - Fotonic, 493–5
 - gyro, 511
 - intrinsic, 492, 497–500
 - magnetic field, 508
 - microbending, 497
 - pH, 500
 - polarimetric, 511–13
 - pressure, 495, 507
 - proximity, 519
 - Raman scattering, 501
 - Rayleigh scattering, 500
 - refractive index, 500
 - rotation, 508–11
 - single mode, 503–8
 - smart structures, 502
 - strain, 498–9, 513
 - temperature, 495, 496, 498–9, 500, 507–8, 513
 - fibers, 359–60, 373–7
 - all plastic, 415–16
 - beam delivery, 280
 - beat length, 418
 - Bragg grating, 420–1
 - cables, 421–3
 - depressed cladding, 392
 - dispersion flattened, 392, 462
 - dispersion shifted, 462
 - elliptical cladding, 420
 - elliptical core, 419
 - fusion splicing, 397
 - graded index, 382–5
 - high birefringence (Hi-Bi), 419
 - in communication systems, 438–72
 - light coupling into, 441–7
 - mid-infrared, 280, 417–18
 - panda, 420
 - plastic-coated silica, 414–15
 - polarization maintaining, 418–20
 - refractive index profile
 - measurement, 410
 - side pit, 420
 - single mode, 385–8
 - step index, 373
 - frequency division multiplexing *see* wavelength division multiplexing
 - indicatrix, 93, 529–30
 - parametric oscillator, 236
 - phonons, 501
 - resonator, 18
 - scanning, 108–10
 - time domain reflectometry (OTDR), 411–12, 500
- ordinary ray, 93, 122
- ordinary refractive index, 93, 122
- output ratio, 403
- outside vapour deposition (OVD), 414
- overlap integral, 446
- packing fraction (fibers), 516, 520
- panda fiber, 420
- parametric
 - amplification, 124
 - lasers, 236–8
 - oscillation, 236–8, 124–6
- paraxial ray, 23
- passive displays, 129, 158
- passive mode locking, 253
- passive Q-switching, 258
- Pauli principle, 42, 57
- penetration depth (electrons), 134
- phase
 - conjugate mirror, 280
 - grating, 421
 - mask, 421
 - shift keying (PSK), 436, 468, 469
 - velocity, 4
- phased arrays (lasers), 220
- phonon
 - acoustic, 501
 - annihilation, 143
 - creation, 143

- optical, 501
- phosphorescence, 129
- phosphors, 129
 - in colour TV, 136
 - in CRT screen, 135
 - in electroluminescence, 138–9
- photocathode, 303–5, 306, 312, 389
 - dark current, 310
 - materials, 303–5
- photoconductive
 - detectors, 314–18
 - bias circuit, 314
 - idealized response, 318
 - noise, 318–19
 - gain, 317
- photodetectors, 293
 - see also* photoconductive detectors, photodiodes, thermal detectors
- photodiode response times, 332–4
 - capacitance effects, 333
 - carrier diffusion, 334
 - carrier drift, 332–3
- photodiodes, 325–41
 - avalanche *see* avalanche photodiode
 - germanium, 456
 - intensified, 343–4
 - metal–semiconductor–metal, 336–7, 484
 - interdigitated structure, 336
 - photoamperic mode, 325, 328
 - photoconductive mode, 325, 327
 - photovoltaic mode, 325, 326–7
 - p–i–n, 329–30
 - quantum well, 86
 - response time, 322–4
 - responsivity, 331
 - Schottky, 82, 334–6
 - silicon, 328–9, 438, 496
 - wavelength response, 330
- photodynamic therapy, 286
- photoelectric effect, 303
- photographic emulsion, 278
- photoluminescence, 131–3
 - definition, 129
 - KCl:Ti (Fig.), 133
 - Stokes shift, 132–3
 - temperature effects, 131
 - use in fluorescent lamps, 133
- photomultiplier, 307–11, 448
 - dynode bias circuit, 307–8
 - dynode structure, 367
 - gain, 307
 - noise in, 310–11
 - photon counting, 352
 - response time, 309
 - secondary electron emission, 307
- photon, 2
 - counting, 352
 - devices, 293
 - wavevector, 142
- photorefractive keratectomy, 286
- photorefractive layer, 278
- phototransistor, 341–2
- p–i–n photodiode, 329–30
 - noise analysis, 448–50
 - use in communications, 438, 448
- pitch
 - GRIN lens, 28
 - liquid crystal, 159
- planar dielectric waveguide, 364–73
 - asymmetric, 472
 - condition for beam propagation, 367
 - modes, 368
 - strongly asymmetric, 473–4
 - field distribution, 474
- Planck's blackbody radiation law, 30
- plaque, 286
- plasma displays, 155–7
- plastic fibers, 415–16
 - attenuation, 416
 - materials, 415
 - Rayleigh scattering, 417
 - use in communications, 388
- plumbicon, 322
- p–n junction, 67–70
 - forward bias, 70–3, 141
 - reverse bias, 73–5, 325
- pneumatic detector, 300
- Pockels
 - effect, 96–7, 532–4
 - modulators (or cell), 97–103, 257–8
 - IO version, 476–7
 - longitudinal, 103
 - transverse, 104
- Poisson statistics, 452
- Poisson's ratio, 498, 507
- polarimetric fiber sensors, 511–13
- polarization, 7
 - circular, 91
 - elliptical, 90–2
 - maintaining fiber, 418–20
 - plane, 7–11
- polarizing beam splitter, 349
- polymethyl methacrylate (PMMA), 415
- population inversion, 175–7
 - in semiconductor lasers, 208–10
 - threshold, 186–9
- positive crystals, 95, 532
- positive liquid crystal material, 160
- potential well, 39–42
- Poynting vector, 34
- principal
 - axis, 94, 528
 - permittivities, 528
 - refractive indices, 528, 529
 - section, 93
- prism, 23, 467
- profile dispersion, 389, 391
- profile parameter, 382, 497
- proximity sensor, 519
- pulse modulation, 429
- pumping, 176–8, 186–90
- pupil (eye), 548
- pyroelectric detectors, 300–2
 - frequency response, 301
- Q-switching, 354–8
 - electro optic, 257–8
 - laser safety, 547
 - passive, 258
 - rotating mirror, 255–6
- quadrature, 504
- quality factor, 192
- quantum
 - confined Stark effect, 86, 117
 - dot, 219
 - efficiency, 141, 536
 - external, 210
 - internal, 210
 - LED, 142, 149 (table), 536
 - photocathode materials (Fig. 1, 305)
 - photoconductors, 317
 - photoemission, 306
 - mechanical tunnelling, 78, 139
 - number, 31, 42, 229–30
 - well, 83–6
 - detector, 332–4
 - modulator, 118–19
 - well laser, 218–19, 486
 - wire, 219
 - yield, 303
- quarter-wave plate, 95
- radiation density, 30, 172
- Raman scattering, 468, 501
- Raman–Nath acousto-optic grating, 114–15
- raster scan, 136
- rate equations (lasers), 187
- Rayleigh
 - criterion, 22
 - scattering, 395–6, 411, 416, 417, 500
 - erbium-doped fibers, 500
 - long wavelength fibers, 417
 - plastic fibers, 416
 - silica fibers, 396, 500
- readout register, 346
- real image (holography), 273
- reciprocity, 509–10
- recombination, 52
 - lifetime, 64
 - mechanisms, 52–3
 - effect on LED response, 152–3, 535–6
 - in semiconductors, 142–4
 - via excitons, 145
 - via impurities, 144–5
- non-radiative, 52
- radiative, 52, 142–4
- reference wave (holography), 273, 278
- reflectance, 8
- surface, 281
- refractive index, 3, 528
 - extraordinary, 93, 529–30

- ordinary, 93, 529
- principal, 528
- relative permittivity, 3, 54, 80, 100
- repeater, 438–9, 457
- resolving power, 22
- resonant transfer, 224
- resonator *see* optical cavity
- response time, 294
- responsivity, 293
- retina, 286, 548
- return-to-zero (RZ) signal, 432
- reverse
 - bias breakdown, 78
 - leakage current, 331
 - saturation current, 141
- Richardson–Dushman equation, 310
- r.m.s. pulse width, 408–9, 541–2
- rods (eye), 548
- R-SEED device, 482
- ruby laser, 199–2
- Sagnac effect, 508
- sampling theorem, 433
- saturable absorber, 253–4, 258
- saturation current, 73
- scanning, 344
- Schottky metal contact, 486
- Schottky photodiode, 82, 334–6, 344
- Schrödinger equation, 39
- second-harmonic generation, 119–24, 124
- semiconductor lasers, 204–23
 - active regions, 205
 - arrays, 219–20
 - buried heterostructure, 213
 - dark-line defects, 217
 - degradation, 216–18
 - distributed feedback, 248, 482
 - GaAs, 205–7
 - heterojunction, 211–18
 - homojunction, 204–7
 - index guiding, 213
 - in optical communications, 438, 446–7, 463
 - kinks, 211, 447
 - power output, 210
 - quantum well, 86, 218–19, 486
 - quaternary compounds, 216
 - short wavelength, 222
 - strained lattice, 222
 - stripe geometry, 211, 217
 - ternary compounds, 211
 - threshold current density, 207–9
 - vertical cavity, 220–2
 - wavelength chirp, 464
- sensitization centres, 318
- sensors (optical fibers) *see* optical fiber sensors
- seven-bar segment display, 163
- shadowmask, 136
- shielding gas (laser welding), 281
- short wavelength semiconductor lasers, 221
- shot noise,
 - avalanche photodiode, 340, 451
 - fiber interferometers, 508
 - photomultiplier, 310
 - p–i–n detector, 448–50
- side pit fiber, 420
- signal wave, 237
- silica phase mask, 421
- silicon-based waveguides, 486–7
- silicon nitride, 222
- silicon photodiode, 328–9
 - electric field distribution, 329
 - p–i–n, 330
 - p–n, 328
 - p–i–n, 330
 - responsivity, 331
 - structure, 328
 - use in communications, 438, 461
- silicon solar cells, 354–5
- single mode fiber, 385–8
 - couplers, 405–6
 - in communications, 457, 462
 - in sensors, 503–8
 - lasers, 244–5
 - mode field diameter, 387
- single rail logic, 484
- skew ray, 374, 385, 516
- skin (laser safety), 548
- slab waveguides, 472
- small signal gain coefficient, 175
- smart structures, 502
- smectic (liquid crystal), 159
- solar cell, 353–5
 - fill factor, 354
 - materials, 354–5
 - ohmic contacts, 354
 - silicon, 354–5
- Soleil–Babinet compensator, 512
- solitons, 470–1
- specific heat, 281
- speckle pattern, 264
- spectral broadening, 183–6
 - collision, 185
 - Doppler, 184
 - homogeneous, 185
 - inhomogeneous, 185, 245
 - natural damping, 185
- spectral response
 - eye, 32, 157
 - photodetectors, 293, 318
- spectrum analyzer (IO), 484
- sphere lens, 26, 444, 445
- spherical aberration, 25
- spherical mirror, 23–4
- spiking oscillation, 199
- spontaneous emission, 170–1, 464
- spontaneous transition lifetime, 170
- sputtering, 475
- S-SEED device, 483
- standing waves, 14, 190
- step index fibers, 373–82
 - intermodal dispersion, 378–82, 457
 - leaky mode, 374
 - meridional ray, 374, 377
 - modes, 374–5
 - numerical aperture, 376
 - refractive index profile, 373
 - skew rays, 374, 377
- stimulated
 - Brillouin scattering, 468
 - emission, 170–1, 502
 - Raman scattering, 468
- Stokes shift, 131, 133, 501
- strained layer, 222
- stripe geometry, 211–13
- stripe waveguides, 475
- sum frequency generation, 125
- superluminescence, 223
- superluminescent diode, 222–3, 510
- superposition principle, 11
- superradiance, 223
- surface reflectance, 281
- surface states, 83
- susceptibility (electric), 119, 527
- taper, 285
- telephone,
 - local networks, 465
 - trunk links, 439
- television, 135–6
- thermal detectors, 293, 296–302, 448
 - bolometer, 298, 300
 - microbolometer, 348
 - pyroelectric, 300–2
 - structure, 297–8
 - theoretical model, 296
 - thermistor, 299
 - thermoelectric, 298
 - time constant, 297
- thermionic emission, 135, 310
- thermistor, 299
- thermoelectric detector, 298
- thermoplastic photoconducting layer, 278
- thin spherical lens, 24–6
- threshold conditions, 181–2
 - current density, 207–10
 - pumping power, 186–9
 - temperature dependence, 214–15
- time constant (detectors), 294
- total
 - internal reflection, 10, 360–4
 - decay in second medium, 364
 - LEDs, 151
 - phase shift, 362
- transatlantic telephone links, 464–5
- transfer function, 408
- transitions
 - atomic dipole, 130
 - forbidden, 130
- transport register, 346
- transverse electric mode fibers, 374
- waveguides, 368

- transverse electromagnetic mode (lasers), 193–5
- transverse magnetic mode
 - fibers, 374
 - waveguides, 368
- transversely excited atmospheric lasers, 232
 - carbon dioxide, 232
 - nitrogen, 233
- traps, 52
 - effect on luminescence, 130–1
 - effect on photoconductor response, 319
 - isoelectronic, 148
- travelling wave modulator, 106
- triglycine sulfate (TGS), 301
- tuning (laser radiation), 235, 237–8, 267
- tunnelling (quantum mechanical), in electroluminescence, 139
- twisted nematic cell, 160–3
- two-beam interference, 14–17
- two-level binary modulation, 432
- Tyndall, 359
- V-groove, 400, 487
- V parameter
 - in optical fibers, 375
 - in planar waveguides, 369, 473
- vacuum level, 62, 303
- vacuum photodiode, 305–6
- valence band, 43, 47
- van Heel, 516
- varactor diode, 75
- vapour axial deposition, 414
- Verdet constant, 111
 - table, 111
- vertical cavity surface-emitting lasers, 220–2
- video displays, 136–7
- vidicon, 321–2
- virtual image (holography), 271
- W fiber, 392
- wave
 - group, 6
 - surface, 4
- wavefront, 4, 366
 - reconstruction, 271–4
- wavefunction, 38–9, 41
- waveguide
 - dispersion, 389, 392
 - quantum well, 86, 486
 - semiconductor laser, 211
 - silicon based, 486–7
- wavelength division multiplexing, 466–8, 469, 486
- wavenumber, 378
- wavevector, 5, 362
- Wheatstone bridge, 299
- white noise, 294
- Wien's displacement law, 29
- wiggler magnet, 238–9
- Wollaston prism, 512
- work function, 62–3, 80, 81, 403
- Xenon fluoride excimer laser, 243
- Young's modulus, 507
- Young's slits, 15, 263
- ZBLAN, 417, 461
- Zener breakdown, 78–9
- zinc selenide (ZnSe), 222

third edition

Optoelectronics

an introduction

John Wilson John Hawkes

The third edition of this best-selling book continues the successful approach adopted by previous editions, and is an ideal text for students on graduate and undergraduate courses, as well as for practising engineers and physicists. Special emphasis is placed on the fundamental principles which underlie the operation of devices and systems.

This is a clearly written, updated text, which incorporates recent interest and developments in optical fibers, sensors and distributed systems. Parametric oscillation and flat-panel displays have been given increased coverage. Topics such as wavelength division multiplexing, optical amplifiers, solitons and coherent systems are introduced, with emphasis on the increased importance of single mode loss fibers and very low loss fibers.

Readers are provided with a thoroughly updated edition that enables them to understand future developments within the subject.

ABOUT THE AUTHORS

Dr Wilson is Dean of Faculty of Engineering, Science and Technology, while Dr Hawkes is an Honorary Research Fellow in the School of Engineering, within the University of Northumbria at Newcastle. They both have wide-ranging experience of teaching optoelectronics and related subjects, and in associated research.

ISBN 0-13-103961-X



9 780131 039612 >